

# Controllable Text Generation

## And Ethical Implications

Shrimai Prabhume

CMU-LTI-21-001

May 2021

Language Technologies Institute  
School of Computer Science  
Carnegie Mellon University  
5000 Forbes Ave., Pittsburgh, PA 15213  
[www.lti.cs.cmu.edu](http://www.lti.cs.cmu.edu)

**Thesis committee:**

Alan W Black (co-chair), Carnegie Mellon University  
Ruslan Salakhutdinov (co-chair), Carnegie Mellon University  
Yulia Tsvetkov, Carnegie Mellon University  
Jason Weston, Facebook AI Research

*Submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy  
In Language and Information Technologies*

©2021, Shrimai Prabhume

*to my late mother, Sushma, for being an embodiment of perseverance and dedication, and for endless love that will last me a lifetime.*



# Abstract

The 21st century is witnessing a major shift in the way people interact with technology and Natural Language Generation (NLG) is playing a central role. Users of smartphones and smart home devices now *expect* their gadgets to be aware of their situation, and to produce natural language outputs in interactions. This thesis identifies three aspects of human communication to make machines sound human-like - style, content and structure. This thesis provides deep learning solutions to controlling these variables in neural text generation. I first outline the various modules which could be manipulated to perform effective controllable text generation. I provide two novel solutions for style transfer – using back-translation technique, and tag and generate approach. I also introduce two new tasks for style transfer and provide datasets for further exploration – political slant transfer and politeness transfer. I establish the task of document grounded generation which leverages information from unstructured documents for the generation process. I introduce two new tasks for document grounded generation – Wikipedia Update generation and Document Grounded Dialogue Response generation. Furthermore, I build two new extensions to pre-trained encoder-decoder models to solve this task. I also design a new elegant solution for the sentence ordering task to learn effective document structures. For all three tasks of style transfer, document grounded generation and sentence order, I add importance to the human evaluation of the models. I introduce new human evaluation measures for understanding the notion of grounding and for understanding the quality of predictions in sentence ordering. At the end, I provide a discussion on the ethical considerations of the applications of controllable text generation. Specifically, I use deontological ethics to evaluate NLP systems and discuss how controllable text generation techniques can be used to make these systems ethical.



# Acknowledgements

I would like to first thank my advisors Alan W Black and Ruslan Salakhutdinov without whom this thesis would not have been possible. Meeting Alan Black was a serendipitous and a pivotal event in my life and I am forever grateful for receiving his insights on not only scientific and technical matters but also about affairs of life. Alan, you are an embodiment of a “guru” from Indian culture; an advisor who has guided me through various walks of life - scientific, technical, professional, spiritual, emotional and personal. I am constantly inspired by your boundless knowledge, tireless spirit for hard work, insatiable thirst to learn new things and generous kindness. Russ is one of the most brilliant researchers of our time and I am extremely grateful that our paths intertwined. Russ, you have been an amazing advisor in every aspect; your vast knowledge, incredible humility, dedication for work, fascinating organization skills and passion for research constantly inspires me. Russ, your mentorship and insights have been central for my growth as a researcher; your flexibility and patience has allowed me to explore interesting problems in domains I would not have ventured otherwise.

I would like to thank my committee members - Yulia Tsvetkov and Jason Weston. Yulia Tsvetkov has provided me with immeasurable support and guidance in the early years of my PhD and has also taught me to write research papers. I would like to thank Jason Weston for detailed feedback on my thesis which greatly improved it.

I would like to thank my collaborators who made projects interesting and fun - Elijah Mayfield, Aman Madaan, Tanmay Parekh, Amrith Setlur, Dirk Hovy, Dheeraj Rajagopal, Brendon Bodlt, and Kangyan Zhou. I would also like to thank Stacey Young for working tirelessly behind the scenes for seamless processes at LTI. I have been fortunate to intern at multiple research labs and my internship work has added a lot of value to this thesis. I would like to thank my intern mentors - Michel Galley, Chris Quirk, Jason Weston, and Kazuma Hashimoto.

I have been very lucky to have a supportive cohort of friends at Pittsburgh. I would like to thank my friends for sharing the burden of my failures and celebrating the joy of my successes - Dheeraj Rajagopal, Vidhisha Balachandran, Shruti Palaskar, Venkat Perumal, Chaitanya Ahuja, Bhavya Balu, Aman Madaan, Bhuwan Dhingra, Rolly Mantri, Priyank Lathwal, Harsh Jhamtani and Sai Krishna Rallabandi.

Finally, I would like to thank my family - my father Laxmikant, and my aunt Purnima for supporting me throughout this endeavor. I would like to thank my dear sister Diksha for being understanding and constantly supporting me through toughest times. Last but definitely not the least, I would like to thank my partner Ankush Das for being my biggest cheerleader throughout the process. Ankush, thank you for unconditional love and incessant support in everyday life, I love you.



# Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Thesis Statement . . . . .	4
1.2 Overview . . . . .	4
<b>2 Controllable Text Generation Techniques</b>	<b>6</b>
2.1 Generation Process . . . . .	7
2.2 External Input . . . . .	9
2.2.1 Arithmetic or Linear Transform . . . . .	9
2.2.2 Stochastic Changes . . . . .	10
2.2.3 Decompose . . . . .	10
2.2.4 External Feedback . . . . .	11
2.3 Sequential Input . . . . .	11
2.3.1 Arithmetic or Linear Transform . . . . .	12
2.4 Generator Operations . . . . .	12
2.4.1 Recurrent Neural Networks . . . . .	13
2.4.2 Transformer . . . . .	14
2.4.3 Pre-trained models . . . . .	15
2.5 Output . . . . .	15
2.5.1 Attention . . . . .	15
2.5.2 External Feedback . . . . .	16
2.5.3 Arithmetic or Linear Transform . . . . .	17
2.6 Training Objective . . . . .	17
2.6.1 General Loss Objectives . . . . .	17
2.6.2 KL Divergence . . . . .	18
2.6.3 Classifier Loss . . . . .	18



---

2.6.4	Task Specific Loss . . . . .	19
2.7	Discussion . . . . .	20
2.8	Conclusion . . . . .	21
<b>3</b>	<b>Style Transfer</b> . . . . .	<b>22</b>
3.1	Tasks and Datasets . . . . .	23
3.1.1	Gender Transfer . . . . .	24
3.1.2	Political Slant Transfer . . . . .	25
3.1.3	Sentiment Modification . . . . .	25
3.1.4	Politeness Transfer . . . . .	26
3.2	Methodology . . . . .	28
3.2.1	Back-translation . . . . .	28
3.2.2	Tag and Generate . . . . .	32
3.3	Experiments . . . . .	36
3.3.1	Style Transfer Accuracy . . . . .	36
3.3.2	Preservation of Meaning . . . . .	38
3.3.3	Fluency . . . . .	40
3.3.4	Manual Inspection . . . . .	41
3.4	Related Work . . . . .	43
3.4.1	Task . . . . .	43
3.4.2	Methodology . . . . .	43
3.5	Conclusion . . . . .	44
<b>4</b>	<b>Document Grounded Generation</b> . . . . .	<b>46</b>
4.1	Tasks and Datasets . . . . .	49
4.1.1	Task Definition . . . . .	49
4.1.2	Wikipedia Update Generation . . . . .	50
4.1.3	Document Grounded Dialog Generation . . . . .	52
4.2	Methodology . . . . .	57
4.2.1	Generative models . . . . .	57
4.2.2	Extractive models . . . . .	59
4.2.3	Pre-trained Encoder-Decoder Models . . . . .	60
4.3	Experiments . . . . .	62
4.3.1	Automated Evaluation . . . . .	62
4.3.2	Human Evaluations . . . . .	65
4.3.3	Manual Inspection . . . . .	68
4.4	Ethical Considerations . . . . .	72
4.5	Related Work . . . . .	74

---

4.5.1	Task	74
4.5.2	Methodology	76
4.6	Conclusion	76
<b>5</b>	<b>Sentence Ordering</b>	<b>78</b>
5.1	Methodology	79
5.1.1	Topological Sort	80
5.1.2	Constraint Learning	80
5.2	Experiments	81
5.2.1	Datasets	81
5.2.2	Baselines	81
5.2.3	Evaluation Metric	82
5.3	Results	83
5.3.1	Discussion	84
5.4	Related Work	85
5.5	Conclusion	86
<b>6</b>	<b>Ethical Considerations</b>	<b>87</b>
6.1	Related Work	89
6.1.1	Ethics	89
6.1.2	Ethics in NLP	90
6.2	Deontological Ethics	91
6.2.1	Generalization Principle	91
6.2.2	Respect for Autonomy	93
6.2.3	Utilitarian Principle	94
6.3	Applying Ethics to NLP systems	94
6.3.1	Question-Answering Systems	95
6.3.2	Detecting Objectionable Content	96
6.3.3	Machine Translation Systems	97
6.3.4	Dialogue Systems	99
6.4	Ethical Decision Making with NLP	100
6.5	Discussion	101
6.6	Conclusion	103
<b>7</b>	<b>Conclusions</b>	<b>104</b>
7.1	Summary of Contributions	104
7.2	Future Directions	106
7.2.1	Broad Directions for Future Work	106

---

7.2.2	Exploring Controllable Text Generation Techniques . . . . .	107
7.2.3	Style Transfer . . . . .	107
7.2.4	Document Grounded Generation . . . . .	109
7.2.5	Ethical Considerations . . . . .	110
7.3	Broader Impact . . . . .	111
7.3.1	Impact beyond NLP . . . . .	111
7.3.2	Impact on Society . . . . .	112
<b>A</b>	<b>Appendix for Style Transfer</b>	<b>116</b>
A.1	Details of Training . . . . .	116
A.2	Additional Results . . . . .	117
A.3	Examples of Generations . . . . .	118
A.4	Preliminary experiments for Future Work . . . . .	122
A.5	Details of Code . . . . .	124
<b>B</b>	<b>Appendix for Document Grounded Generation</b>	<b>125</b>
B.1	Details of Training . . . . .	125
B.2	Additional Dataset Details . . . . .	127
B.3	Details of Code . . . . .	139
<b>C</b>	<b>Appendix for Sentence Ordering</b>	<b>140</b>
C.1	Details of Training . . . . .	140
C.2	Examples of Sentence Order Predictions . . . . .	140
C.3	Details of Code . . . . .	143
	<b>Bibliography</b>	<b>144</b>

# Chapter 1

## Introduction

*“The common misconception is that language has to do with words and what they mean. It doesn’t. It has to do with people and what they mean.”*

*Herb Clark and Michael Schober, 1992*

One of the important goals of artificial intelligence (AI) is to model and simulate human intelligence. Modeling human interactions is a sub-goal in the path of answering the larger question on human intelligence. Natural Language Generation (NLG) is an important aspect of modeling human communications. NLG by definition focuses on producing human languages<sup>1</sup> from non-linguistic machine representations of knowledge. The process of generating language poses an interesting question of how information is best communicated between a machine and a human.

This thesis is inspired by the research question *“Should machines reflect the way humans interact in society?”*. I have identified three aspects of human communication that I am interested in using for generation: (1) Style (2) Content and (3) Structure. Style is used in human communication to convey specific goals effectively and also to define social identities. All human communications carry some degree of information in them, which I call content. One way communications or documentations such as blogs, memos, reports etc. also enclose relevant information. The ordering of information in these communications is structure and each of these communication goals requires different structures to achieve the desired effect.

Most human practices display style (Coupland, 2007). For example, fashionable style is reflected in the choice of clothes and accessories we wear, architectural style is exhibited in the choice of raw materials used, color, design plans etc. of the construction, culinary style is demonstrated in the raw materials, size and color of crockery, etc. Similarly, linguistic style is expressed in the choice of words or phrases as well syntactic structures used to convey a

---

<sup>1</sup>Although the philosophy and techniques mentioned in this thesis are applicable to any natural language, I focus only on English (#BenderRule).

piece of information. Note that ‘style’ in computational linguistics is a loaded term and I don’t partake in disambiguating its usage. I define style as a group of natural language sentences that belong to a particular class or label. I focus on controlling the neural generation process to adhere to a specific style. In particular, I propose the novel approach of using neural back-translation for building a hidden representation that has reduced stylistic elements but is grounded in semantic meaning to the input sentence. I finally use an adversarial training objective to ensure that the generation complies with the target style.

Human communication by definition is a process by which individuals exchange information and influence one another through a common system of symbols and signs (Higgins and Semin, 2001). This behavior is however not mirrored in natural language generation systems. Typically, models hallucinate information to be generated as they are not conditioned on any external source of knowledge. Generating natural language from schematized or structured data such as database records, slot-value pair, Wikipedia Infobox etc. has been explored in prior work (Mei et al., 2016; Wen et al., 2015; Lebret et al., 2016). A lot of information resides in unstructured format in the form of books, Encyclopedias, news articles, Wikipedia articles etc. I focus on leveraging this information to guide the generation process to include relevant pieces in the generated text. I propose various neural models to incorporate both context and an external source of information into the generation step.

Human beings effortlessly produce complicated pieces of text that are well connected and appropriately ordered (Hovy, 1993). Most effective human communication is not in the form of randomly ordered information but it is well planned and structured. In spite of the recent advances in natural language processing (NLP), NLG systems have not gained the ability to plan and organize multiple sentences. I focus on solving the sentence ordering sub-task which involves ordering the information in a document. Sentence ordering is the task of arranging the sentences of a given text in the correct order. In particular, I pose this task as a constraint solving problem and leverage rich sentence representations provided by pre-trained language models to design these constraints.

Reiter and Dale (2000) detail seven sub-tasks which are conceptually distinct to describe the generation process. These sub-tasks can be modeled separately or in some cases they may interleave. In (Reiter and Dale, 2000), these seven sub-tasks are primarily characterized as content or structure tasks. Contrary to this characterization, I connect the style, content and structure aspects of this thesis to the different sub-tasks in (Reiter and Dale, 2000). The seven sub-tasks are: (1) *Content Determination* is the sub-task of deciding what information needs to be communicated in the generated piece of text. (2) *Document Structuring* is the sub-task of grouping similar content together and then deciding the relations between the groups to generate a coherent structured text. (3) *Lexicalization* is the sub-task of choosing specific set of phrases or other linguistic features such as syntactic constructs to express the selected content in the desired manner. (4) *Referring Expression Generation* is involved with selecting the desired expressions to be used to refer to entities. (5) *Aggregation* is concerned with mapping document structures onto linguistic structures such as sentences and paragraphs. This sub-task can also

decide the ordering of information that has to be generated. (6) *Linguistic Realisation* is the sub-task of converting abstract representations of sentences into the real text. (7) *Structure Realisation* is the sub-task of converting abstract structures such as paragraphs and sections into mark-up symbols and segments understood by humans.

Style is related to the *lexicalization* sub-task and I control the generation process by selecting the desired phrases or other linguistic resources. Content is the *content determination* sub-task and I guide the generation process with explicit information that is needed in the generated text. I focus on understanding document structures and hence appeal to the *document structuring* sub-task and provide an elegant solution for ordering of sentences for the *aggregation* sub-task in my exploration of structure. Note that the *linguistic realisation* sub-task is already solved by sequence-to-sequence frameworks which generate sentence from a hidden representation of it. The sub-task of deciding document structure boundaries in *structure realisation* and *referring expression generation* is left for future work.

At a minimum, controlling these three aspects of communication can be used for tasks such as:

- Dialogue systems - controlling the persona of the system, various aspects of the response such formality, authority etc., and grounding conversation on unstructured content.
- Story generation - introducing NLG into audience-appropriate narrative texts, generating stories from given plots or events.
- Report generation - pulling disparate source documents into a coherent unified whole, which can use a shared set of sources to generate a variety of genres:
  - News articles covering current events with historical context.
  - Wiki articles summarizing a topic's evolution over time.
  - Scientific article summaries highlighting key findings on a topic.

Human communications also sometimes carry debatable features such as usage of swear words and obscenity in language (McEnery, 2005), or using the power of language to target minority groups to project social biases and reinforce stereotypes on people (Fiske, 1993). Providing fine grained control on style, content and structure in generated text runs the risk of generating language which has undesirable consequences such as spewing hate or targeting groups to promote violence or social disorder. In the last part of my thesis, I open the discussion on the ethical considerations of controllable text generation. While mirroring the style, content and structure aspects of human communication, it is also important to think about the scenarios when we don't want the machines to reflect human interactions. On the other hand, I also explore how controllable text generation can be used to make systems ethical.

## 1.1 Thesis Statement

*Controlling style, content and structure leads to human-like generations which should be used with ethical considerations.*

The central goal of this thesis is to integrate aspects of human communication in natural language generation systems to make them sound human-like. To fulfill this goal, I focus on controlling the style, content and structure aspects of natural language generation. Within each aspect, this thesis contributes new tasks as well as new models. For making modeling contributions, it is important to understand the landscape of different architectures from prior work. Hence, I start by presenting a new schema for controllable text generation which collates knowledge about various architectures from prior work in different domains and tasks. For controlling the *style* aspect, I specifically focus on the task of style transfer and introduce two new style transfer tasks and two new approaches to explore and understand this aspect. For controlling the *content* aspect, I design and introduce the new line of research on document grounded generation. Specifically, I introduce two concrete tasks: *Wikipedia Update Generation* and *Document Grounded Dialogue Generation*. I also contribute towards two new extensions of pre-trained encoder decoder models. For controlling the *structure* aspect, I focus on the sentence ordering sub-task and contribute a new framing of this task. For each of the three aspects, this thesis pays significant attention to human evaluation. I point the limitations of automatic metrics and the need for better evaluation. Ethical implications of controllable text generation systems could also be one way of evaluating NLP systems. For any NLP system that interacts with humans, it is important to understand the ethical considerations of that technology. Hence, in the last part of this thesis, I explore not only the ethical considerations of controllable text generation but also how controllable text generation can be used to make NLP systems ethical. This thesis provides a range of answers to new research questions as well as new statistical models. The practical contributions of the thesis are new tools and new large datasets.

## 1.2 Overview

**Controllable Text Generation Techniques (Chapter 2):** I start by providing the necessary technical background for understanding this thesis. In this chapter, I connect the different works in controllable text generation and collate the knowledge about the similarities of these tasks and techniques. I organize the prior work and propose a new schema which contains five modules that can be changed to control the generation process - external input module, sequential input module, generator operations module, output module and the training objective module. I lay grounds to the different theories of representing control vectors and incorporating them into the generation process as well as provide a qualitative assessment of these techniques.

**Style Transfer (Chapter 3):** This chapter talks about the importance of *style* aspect in human communication. I focus on the *style transfer* task in this thesis. I describe two novel approaches to perform style transfer in non-parallel data: (1) back-translation for style transfer, and (2) tag and generate approach. I introduce two new tasks for exploring style transfer: (1) political slant transfer, and (2) politeness transfer. I outline and also provide insights in both automatic and human evaluations for three dimensions of accessing style transfer methods: style transfer accuracy, preservation of meaning and fluency.

**Document Grounded Generation (Chapter 4):** This chapter provides an overview on the different tasks for content grounded generation. It focuses on the content aspect of human communication. A formal definition to the new task of document grounded generation is provided. Specifically, I propose two new tasks for grounded generation in two different domains. First is Wikipedia edit generation task which is concerned with generating a Wikipedia update given an external news article and the Wikipedia article context. Second is dialogue response generation which involves generating a response based on the knowledge from an external source and the current dialogue history. Experiments are performed with both generative as well as extractive models to solve these tasks. Due to the recent success of pre-trained models, a strong baseline of pre-trained encoder-decoder is provided. Additionally, two new extensions to the pre-trained models are also proposed. Two new human evaluations are proposed for this task and adopt absolute human evaluation from prior work. A comprehensive manual inspection of the generated outputs is showcased in this chapter.

**Sentence Ordering (Chapter 5):** In this chapter, I provide an overview of the different techniques used to capture document structures. I particularly focus on the sub-task of sentence ordering and propose a new framing of this problem as a constraint solving task. I also introduce a new model based on the new design of the problem. I suggest a new human evaluation for this task which analyzes the human choices for predicted orders in comparison to the reference orders.

**Ethical Considerations (Chapter 6):** With this chapter I start the discussion on the ethical considerations of controllable text generation. I give an overview of the various ethical issues pertaining to NLP and the need to discuss these issues. I extensively discuss the current literature on ethics and the missing parts. I also provide a summary of three principles of ethical science - *generalization principle*, *respect for autonomy* and *utilitarian principle*. Additionally, I present four case studies that discuss how these principles can be applied to popular NLP systems. For each of the case studies, I identify practical NLP techniques that can be used to improve those systems from an ethical stand point. Further, I also provide a comprehensive discussion on the limitations of these approaches and how controllable text generation can be used to make these systems ethical.



## Chapter 2

# Controllable Text Generation Techniques

Controllable text generation is the task of generating natural sentences whose attributes can be controlled. The attributes to control can range from being stylistic such politeness, sentiment, formality, etc.; demographic attributes of the person writing the text such as gender, age, etc.; content such as information, keywords, entities, etc.; ordering of information, events, like plot summaries etc. Controlling various attributes of text generation has manifold applications. For instance in dialogue response generation task, work has been done in controlling persona (Zhang et al., 2018; Li et al., 2016b), controlling various aspects of the response such as politeness (Niu and Bansal, 2018a), formality, authority etc, grounding the responses in external source of information (Zhou et al., 2018; Dinan et al., 2018; Ghazvininejad et al., 2018), and controlling topic sequence (Tang et al., 2019; Prabhumoye et al., 2020b). Another application is story generation where you can control the ending (Peng et al., 2018), the persona (Chandu et al., 2019b), the plot (Yao et al., 2019), and the topic sequence (Huang et al., 2019b). Controllable text generation is also used to modulate the formality and politeness of emails (Madaan et al., 2020b). Report generation can be controlled by pulling disparate source documents into a coherent unified whole, which can use a shared set of sources such as Wikipedia article generation (Liu et al., 2018; Prabhumoye et al., 2019b).

Although there is a large body of prior work in controllable text generation, there is no unifying theme. Each work addresses a specific task in a specific context. In this chapter, we outline a new schema which connects prior work and provides an insight into various aspects of controllable text generation. The schema contains five modules that cover the overall generation pipeline and provide an understanding of the effect of each component on the generation process. Prior work has focused on specific parts of the schema that we outline here and we provide insights into their similarities. We provide an overview of these modules and also present an exploration of the various techniques used to control and update each of these modules.

This schema provides an insight into the contributions of the various modules for controllable text generation. The main advantage of this schema is that it can be used with any algorithmic paradigm like sequence-to-sequence, adversarial methods, reinforcement learning, etc. The schema can also be used with non-autoregressive algorithms which may generate text using graphical structures like trees (Welleck et al., 2019; Guo et al., 2019). In this chapter, we focus on how this schema can be used to describe controllable text generation focusing particularly on the use of autoregressive models. This work paves way to designing new architectures based on our schema. This can be done by identifying promising techniques for each module and then combining them. Our schema can also be potentially used for applying these techniques on new tasks of similar nature. It also provides a standardized framework to position and compare new architectures with existing techniques.

The prior work on unifying text generation models has mostly focused on building efficient tool-kits and modular views of generation. For instance, (Reiter and Dale, 2000) details seven sub-tasks which are conceptually distinct to describe the generation process. These sub-tasks can be modeled separately or in some cases they may interleave. In (Reiter and Dale, 2000), these seven sub-tasks are primarily characterized as content or structure tasks. Note that Reiter and Dale (2000) is not specific to neural text generation. Our work focuses specifically on controlling attributes in neural text generation process. We don't divide the generation pipeline into several sub-tasks but we divide the neural text generation process into modules all of which are required for generation. In (Hu et al., 2019), the focus is on building a toolkit for various text generation tasks based on the three properties of versatility, modularity and extensibility. This work enlists few model architectures and learning paradigms for various text generation tasks. We focus only on the generation process of controllable text generation tasks. We specifically detail the inputs, outputs and operations of the generation process. We do not provide any specific examples of architectures but provide an overview of the basic underlying modules which can be used with any learning paradigm. Xie (2017) provides a practical guide to the neural generation process describing it in terms of initialization, optimization, regularization and decoding strategies. Our work on the other hand does not delve into the implementation details of the generation pipeline but provides an overall schema for understanding of the various components involved.

## 2.1 Generation Process

Most of the controllable text generation tasks can be framed as conditional language generation tasks. Given a corpus of tokens  $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_T)$ , the task of language modeling is to estimate the joint probability  $P(\mathbf{U})$ , which is often auto-regressively factorized as  $P(\mathbf{U}) = \prod_t^T P(\mathbf{u}_t | \mathbf{u}_{<t})$ . For this thesis, we consider conditional language model which has an input or a *source* sequence  $\mathbf{U}$  and an output or *target* sequence  $\mathbf{Y}$  to be generated. In this case we model the probability of the *target* sequence conditioned on the *source* sequence given by  $P(\mathbf{Y} | \mathbf{U}) = \prod_t^T P(y_t | \mathbf{U}, y_{<t})$ . Sequence-to-sequence models which refers to the broader

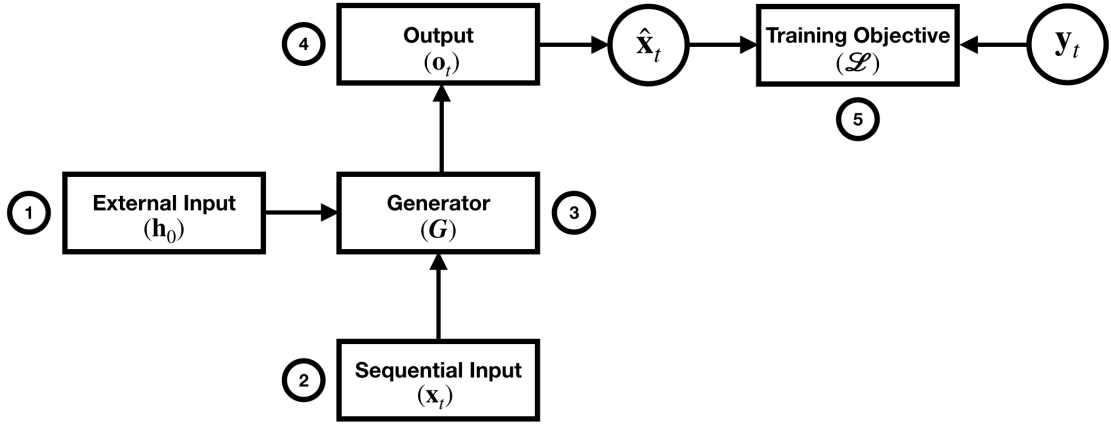


FIGURE 2.1: Modules that control the generation process

class of models that map one sequence to another, are generally used to build conditional language models. The representation of the probability  $P(\mathbf{U})$  of the *source* sequence given by a neural model is denoted by  $\mathbf{h}_e$ . The initialization of the standard generation process  $\mathbf{h}_0$  is equal to  $\mathbf{h}_e$ . The generation of the target tokens of the sequence  $\mathbf{Y}$  unfolds as a time series where each token  $y_t$  is generated at a time step  $t$ .

At a given time step  $t$ , a generative model performs some set of operations  $G$  by taking in as input the previous hidden state  $\mathbf{h}_{t-1}$  and the input  $\mathbf{x}_t$ . Note that the hidden state  $\mathbf{h}_{t-1}$  represents the probability of the tokens generated up to time step  $t$  as well as the *source* sequence  $\mathbf{U}$ .  $\mathbf{x}_t$  is the word embedding of the token  $y_{t-1}$ . The generator produces an output hidden state  $\mathbf{h}_t$  at the current time step. In the standard generation process the output state  $\mathbf{o}_t$  is equal to the hidden state  $\mathbf{h}_t$ .  $\mathbf{o}_t$  is projected to the vocabulary space using a linear transform given by  $\mathbf{W}_o \mathbf{o}_t + \mathbf{b}_o$  which is used to predict token  $\hat{x}_t$  using decoding strategies. Typically an  $\text{argmax}$  function is used as a decoding strategy which means that the token with the highest probability at the current time step is predicted. The ground truth token to be generated is denoted by  $y_t$  and a loss  $\mathcal{L}$  is computed by comparing  $y_t$  to  $\hat{x}_t$ .

**Overview:** In the remainder of the chapter, we provide an overview of the schema which contains five modules that can be used for controlling the generation process (shown in Figure 2.1):

1. **External Input** module is responsible for the initialization  $\mathbf{h}_0$ , of the generation process (§2.2).
2. **Sequential Input** module is the input  $\mathbf{x}_t$  at each time step of the generation (§2.3).
3. **Generator Operations** module performs consistent operations or calculations on all the input at each time step (§2.1).
4. **Output** module is the output  $\mathbf{o}_t$  which is further projected on to the vocabulary space to predict the token  $\hat{x}_t$  at each time step (§2.5).

5. **Training Objective** module takes care of the loss functions used for training the generator (§2.6).

This work is done in collaboration with Alan W Black and Ruslan Salakhutdinov (Prabhumoye et al., 2020a).

## 2.2 External Input

In this section we discuss the different techniques which can be used to control the generation process by controlling  $\mathbf{h}_0$ . This is marked as position 1 in Figure 2.1.

### 2.2.1 Arithmetic or Linear Transform

One of the easiest ways to control the generation is to concatenate a control vector  $\mathbf{s}$  to output of the encoder. Let the output of the encoder be  $\mathbf{h}_e$  (described in §2.1), then the initialization of the decoder  $\mathbf{h}_0$  will be  $[\mathbf{h}_e; \mathbf{s}]$ , where  $[a; b]$  denotes concatenation. Here, the control vector  $\mathbf{s}$  would provide the generator with a strong signal to guide the generation process.

Fu et al. (2018) use this technique to control the style representation for their generator. The encoder builds representation that is devoid of the style and only retains content. The control vector for style is then concatenated to the encoder representation to initialize the decoder. This technique is commonly used in (Ghazvininejad et al., 2018; Zhou et al., 2018) to concatenate information from external sources to dialogue context to generate dialogue responses. Chandu et al. (2019b) concatenate personality representation  $\mathcal{P}$  derived from a separate corpus to generate visual stories. They also experiment with a simple arithmetic operation on  $\mathbf{h}_e$  given by  $\mathbf{h}_0 = \mathbf{h}_e - \mathcal{S} + \mathcal{P}$  to get the initialization of the generator (here  $\mathcal{S}$  denotes the average representation of the story). They observed that while concatenation technique is better at preserving the meaning of the generated story, the arithmetic operation provides a better signal of the personality for the generation process.

Hoang et al. (2016) uses both the concatenation technique as well as performs a linear transform of  $\mathbf{s}$  to obtain  $\mathbf{h}_0$  for language modeling task. The control vectors in this case represents meta data such as key-words, topics etc. In case of the linear transform  $\mathbf{h}_0 = \tanh(\mathbf{W}_1 \mathbf{h}_e + \mathbf{W}_2 \mathbf{s} + \mathbf{b})$ . The paper also explores adding the control vector to the encoder representation ( $\mathbf{h}_0 = \mathbf{h}_e + \mathbf{s}$ ).

In case of addition, the resulting  $\mathbf{h}_0$  would be averaged representation of the input representation  $\mathbf{h}_e$  and  $\mathbf{s}$ . Information could be lost in this case as control is not explicit. In case of concatenation, if the size of the control vector  $\mathbf{s}$  is too small compared to the context vector  $\mathbf{h}_e$ , then  $\mathbf{s}$  is over-shadowed by  $\mathbf{h}_e$  and the generator will not be able to pay attention to  $\mathbf{s}$ .

Hence it is important to choose comparable dimensions for these two vectors. But this increases the size of model considerable and could be quite costly. Linear transform avoids these issues and performs better than the other two techniques for Hoang et al. (2016).

### 2.2.2 Stochastic Changes

Kingma and Welling (2014) introduce the variational auto-encoder (VAE), where you can stochastically draw a continuous latent variable  $\mathbf{z}$  from a Gaussian distribution. The initialization of the generator  $\mathbf{h}_0$  is based on this latent variable which is drawn. Bowman et al. (2016) use this concept for generating sentences from this continuous latent representation. This process of changing the encoder state  $\mathbf{h}_e$  is can only be used with Kullback-Leibler (KL) Divergence training objective described in (§2.6).

In (Wang et al., 2019b), VAE is used to guide the generation process with topics of a document. A gaussian mixture model is used to incorporate topics into latent variables. In (Xu et al., 2019), VAE is used to control for sentiment attribute in style transfer task by constraining the posterior mean to a learned probability simplex.

Such a design of controllable text generation works when the control attributes can be represented as latent variables for example style, topics, strategies etc. This design will not work for content grounded text generation tasks where specific information, keywords or entities have to guide the generation process.

### 2.2.3 Decompose

You can decompose the encoder representation  $\mathbf{h}_e$  into multiple subspaces, each of which signifies a different attribute you would like to control. Liu and Lapata (2018) split the encoder representation  $\mathbf{h}_e$  into two components, one which represents the structure in the document and the other represents the semantic information. This formulation was used by (Balachandran et al., 2020) for controlling structure in abstractive summarization. This work performs the split with respect to the dimensions of  $\mathbf{h}_e$ . The method forces the first  $n$  dimensions of  $\mathbf{h}_e$  to capture meaning and the latter to capture structure. Balachandran et al. (2020) also show quantitative and qualitative analysis on the types of structures of documents learnt by this technique.

Romanov et al. (2019) decompose the encoder representation  $\mathbf{h}_e$  into a form vector  $\mathbf{f}$  and a meaning vector  $\mathbf{m}$ . During the training phase, a *discriminator* enforces  $\mathbf{m}$  to not carry any information about the form using an adversarial loss and a *motivator* is used for a motivational loss that encourages  $\mathbf{f}$  to carry the information about the form. The generation process can then be guided to adhere to the desired target form. As opposed to splitting  $\mathbf{h}_e$  with respect to dimensions, this work learns subspaces  $\mathbf{W}_m$  and  $\mathbf{W}_f$  given by  $\mathbf{m} = \tanh(\mathbf{W}_m \mathbf{h}_e + \mathbf{b}_m)$  and  $\mathbf{f} = \tanh(\mathbf{W}_f \mathbf{h}_e + \mathbf{b}_f)$  respectively. When  $\mathbf{h}_e$  is projected on  $\mathbf{W}_m$ , we get the meaning vector

$\mathbf{m}$  and similarly when it is projected on  $\mathbf{W}_f$  we get the form vector  $\mathbf{f}$ . This work shows qualitatively how  $\mathbf{m}$  and  $\mathbf{f}$  are learnt in the subspaces using t-SNE plots. It also shows quantitatively the use of  $\mathbf{m}$  and  $\mathbf{f}$  in downstream paraphrase detection tasks. This is an excellent method in building interpretable representations for control attributes. Although, the effectiveness of this technique is not yet proven in the style transfer task or the abstractive summarization task. In both the above mentioned works, the models learn interpretable representations of control attributes but were not able to beat state of the art methods in their respective tasks. It is also worth noting that learning good decomposed vectors is especially hard when no supervision is provided on what the decomposed components are supposed to learn.

This technique works well when the representation space of the input  $\mathbf{x}$  can be decomposed into subspaces which represent different control attributes. This means that the input  $\mathbf{x}$  needs to contain signal of the control attributes. It will not work when the control attributes need to be externally provided. For example in case of content grounded generation tasks described in (Prabhumoye et al., 2019b; Dinan et al., 2018; Zhou et al., 2018), the input may not necessarily contain the content that needs to be generated. A separate input of the content to be generated is provided in these cases.

#### 2.2.4 External Feedback

A regularizer is often used to control the external input  $\mathbf{h}_0$  to the generator. In many cases, an adversarial loss to manipulate the latent space is used as an external feedback mechanism. This essentially controls the latent space of the encoder which is eventually provided as an initialization to the generator. In (Fu et al., 2018), a multi-layer perceptron (MLP) is used for predicting the style labels from  $\mathbf{h}_0$ . Similarly, the adversarial loss is also used in (Wang et al., 2019a) to control the latent representation  $\mathbf{h}_0$  for style attributes. In (Romanov et al., 2019), an adversarial loss is used to ensure that the meaning representation  $\mathbf{m}$  does not carry any style signals. The adversarial loss is obtained by training a discriminator which takes as input a representation  $\mathbf{m}$  and tells if it carries the target style signal. Similarly, this work also employs a motivator loss which is the opposite of the adversarial loss to ensure that the style representation  $\mathbf{f}$  actually does carry the stylistic information. John et al. (2019) use multiple losses to control the style and content information represented in  $\mathbf{h}_0$ .

The discriminator which provides external feedback has to be jointly trained with the generator. This technique can be useful with the decompose technique to ensure that the decomposed sub-spaces represent the desired control attributes.

### 2.3 Sequential Input

In this section we discuss the different techniques which can be used to control the generation process by controlling the sequential input  $\mathbf{x}_t$  to the decoder at each time step. This is marked

as position 2 in Figure 2.1.

### 2.3.1 Arithmetic or Linear Transform

Similar to changing the initialization, we can change the input to the decoder by concatenating the information at each time step with some additional control vector  $\mathbf{s}$ . Typically, teacher forcing method (Williams and Zipser, 1989) is used to train the generator. At time step  $t$ , the generator takes as input the word embedding  $\mathbf{x}_t$  of the word that was predicted at step  $t - 1$  and predicts the word to be generated  $\mathbf{y}_t$  at the current time step. Note that  $\mathbf{x}_t = \mathbf{y}_{t-1}$ . The input  $\mathbf{x}_t$  can be concatenated with  $\mathbf{s}$  at each time step to control the generation process. Hence,  $\tilde{\mathbf{x}}_t = [\mathbf{x}_t; \mathbf{s}]$ .

Noraset et al. (2017), use this technique in the task of definition modeling. They concatenate word embedding vector  $\mathbf{s}$  of the word to be defined at each time step of the definition generation process. Unfortunately, for this task, this technique has not proved to be effective compared to other techniques of controlling the generation. Zhou et al. (2018) concatenate the hidden representation of the external source of information  $\mathbf{s}$  to each time step of dialogue response generation. Similarly, Prabhumoye et al. (2019b) also concatenate the hidden representation of the external source of information  $\mathbf{s}$  to each time step of Wikipedia update generation process. In this work as well, this results of this technique were not as impressive as simple concatenating the control context to the input of the encoder. Harrison et al. (2019) concatenate a side constraint  $\mathbf{s}$  which represents style and personality into the generation process. For this task of generating language from meaning representations with stylistic variation, this method performed better than conditioning the encoder with side constraint in terms of BLEU metric. Chandu et al. (2019b) also concatenate the personality representation  $\mathcal{P}$  at each time step of the story generation process. This is used to control the personality of the visual stories. In addition to concatenation, this work proposes to modify the sequential input as  $\tilde{\mathbf{x}}_t = \mathbf{x}_t - \mathcal{S} + \mathcal{P}$  (here  $\mathcal{S}$  denotes the average representation of the story and  $\mathcal{P}$  denotes the representation of the personality). The latter technique is better at generating personality conditioned stories than the concatenation technique. Neither of these techniques prove to be conclusively better than making similar changes to the external input module (§2.2). Note that in this technique, changes are made directly to the input of generation and not the context which is the case with external input. Also, most of the prior work has focused on recurrent neural network and its variants for making such changes. It could be interesting to see such changes made to transformers (Vaswani et al., 2017).

## 2.4 Generator Operations

This module takes in the external input  $\mathbf{h}_0$ , the sequential input  $\mathbf{x}_t$  at time step  $t$  and performs computation to return an output  $\mathbf{o}_t$ . Different set of operations can be performed to compute

$\mathbf{o}_t$  which are enlisted below. You can also decide to change the operations based on the control vector  $\mathbf{s}$  to compute  $\mathbf{o}_t$ . This is shown as position 3 in Figure 2.1.

### 2.4.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are designed to model sequential information. RNNs perform the same operations for every element of a sequence, with the output depending on previous computations. This recurrence serves as a form of memory. It allows contextual information to flow through the network so that relevant outputs from previous time steps can be applied to network operations at the current time step. Theoretically, RNNs can make use of information in arbitrarily long sequences, but empirically, they are limited to looking back only a few steps.

The Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) units are a type of RNNs that have additional ‘memory cell’ apart from standard units of basic RNNs. The memory cell can maintain information in memory for long periods of time. A set of gates is used to control when information enters the memory, when it’s output, and when it’s forgotten. This architecture lets them learn longer-term dependencies. The vanishing gradient problem of RNNs is resolved here. Gated Recurrent Units (GRUs) (Cho et al., 2014) are similar to LSTMs, but use a simplified structure designed to adaptively capture dependencies of different time scales. They also use a set of gates to control the flow of information, but they don’t use separate memory cells, and they use fewer gates.

The computations of the RNN or its variants can be modified to account for the control attribute. Additional gates can be added or the control attribute can be provided as an additional input to the standard gates of RNNs. Gan et al. (2017) propose a variant of the LSTM model, named factored LSTM, which controls style representation in image caption task. The parameters of the LSTM module which are responsible to transform the input  $\mathbf{x}_t$  are factored into three components  $\mathbf{U}$ ,  $\mathbf{S}$  and  $\mathbf{V}$ . The operations of the input ( $\mathbf{i}_t$ ), forget ( $\mathbf{f}_t$ ) and output gate ( $\mathbf{o}_t$ ) are given by:

$$\begin{aligned}\mathbf{i}_t &= \text{sigmoid}(\mathbf{U}_{ix}\mathbf{S}_{ix}\mathbf{V}_{ix}\mathbf{x}_t + \mathbf{W}_{ih}\mathbf{h}_{t-1}) \\ \mathbf{f}_t &= \text{sigmoid}(\mathbf{U}_{fx}\mathbf{S}_{fx}\mathbf{V}_{fx}\mathbf{x}_t + \mathbf{W}_{fh}\mathbf{h}_{t-1}) \\ \mathbf{o}_t &= \text{sigmoid}(\mathbf{U}_{ox}\mathbf{S}_{ox}\mathbf{V}_{ox}\mathbf{x}_t + \mathbf{W}_{oh}\mathbf{h}_{t-1}) \\ \tilde{\mathbf{c}}_t &= \text{tanh}(\mathbf{U}_{cx}\mathbf{S}_{cx}\mathbf{V}_{cx}\mathbf{x}_t + \mathbf{W}_{ch}\mathbf{h}_{t-1})\end{aligned}$$

Particularly, the matrix set  $\{\mathbf{S}\}$  is specific to each style in the task and is responsible to capture the underlying style features in the data.

In (Kiddon et al., 2016), the GRU unit is modified to accommodate extra inputs - goal  $\mathbf{g}$  and agenda items  $E_t^{new}$  in the recipe generation task. The operation of the new component  $\tilde{\mathbf{h}}_t$  is



given by:

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_h \mathbf{x}_t + \mathbf{r}_t \odot \mathbf{U}_h \mathbf{h}_{t-1} + \mathbf{s}_t \odot \mathbf{Y} \mathbf{g} + \mathbf{q}_t \odot (\mathbf{1}_L^T \mathbf{Z} \mathbf{E}_t^{new})^T)$$

where  $\mathbf{s}_t$  is a goal select gate and  $\mathbf{q}_t$  is a item select gate. With this modification, the generation process is controlled for the items to be generation in the recipe and the goal.

Wen et al. (2015) adapt the LSTM to control the dialogue act information in the generation process. The operation to compute the cell value  $\mathbf{c}_t$  is given by:

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t + \tanh(\mathbf{W}_d \mathbf{d}_t)$$

The dialogue act representation  $\mathbf{d}_t$  is build using another LSTM cell.

RNNs, LSTMs and GRUs are commonly used to model sequence-to-sequence controllable text generation tasks (Prabhumoye et al., 2019b; Rao and Tetreault, 2018; See et al., 2017; Zhou et al., 2018; Fu et al., 2018).

## 2.4.2 Transformer

Transformers are proposed by (Vaswani et al., 2017) and they rely on attention mechanism to draw global dependencies between input and output. The Transformer uses stacked self-attention and point-wise, fully connected layers for both the encoder and decoder. The encoder stacks  $N$  identical layers, each of which has two sub-layers. The first sub-layer is a multi-head self-attention mechanism (described in §2.5), and the second sub-layer is a positionwise fully connected feed-forward network. Each sub-layer uses residual connections around each of the sub-layers, followed by layer normalization. The decoder has an additional third sub-layer, which performs multi-head attention over the output of the encoder stack. This is known as cross-attention layer.

Each of the multi-heads are of the form:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= [\mathbf{H}_1; \dots; \mathbf{H}_m] \mathbf{W}^o, \\ \mathbf{H}_j &= \text{Attention}(Q \mathbf{W}_j^Q, K \mathbf{W}_j^K, V \mathbf{W}_j^V). \end{aligned}$$

The multi-head function receives three inputs - a query  $Q$ , key  $K$  and value  $V$ .  $\mathbf{W}^o$  is an output projection of the concatenated outputs of the attention heads. Each  $\mathbf{H}_j$  is the output of a single attention head and  $\mathbf{W}_j^Q$ ,  $\mathbf{W}_j^K$  and  $\mathbf{W}_j^V$  are head-specific projections for  $Q$ ,  $K$ , and  $V$ , respectively.

Each decoder layer follows the following sequence of functions:

$$\begin{aligned} \mathbf{h} &= \mathcal{F}(\text{self-attention}(\mathbf{h}_x, \mathbf{h}_x, \mathbf{h}_x)), \\ \mathbf{h} &= \mathcal{F}(\text{cross-attention}(\mathbf{h}, \mathbf{h}_u, \mathbf{h}_u)), \\ \mathbf{h} &= \mathcal{F}(\text{FFN}(\mathbf{h})), \end{aligned}$$

where  $\mathcal{F}(\mathbf{h})$  is a sequence of  $\text{LayerNorm}(\text{residual} + \text{dropout}(\mathbf{h}))$ , followed by  $\text{residual} = \mathbf{h}$ ;  $\mathbf{h}_x$  is the embedding of the decoder input up to the current time step,  $\mathbf{h}_u$  is the representation of the *source* sequence  $\mathbf{U}$ .

### 2.4.3 Pre-trained models

Recently pre-trained conditional language models are used for text generation like GPT (Radford et al., 2018), GPT2 (Radford et al.), XLNet (Yang et al., 2019), etc. Several works have fine-tuned the pre-trained models for downstream controllable text generation tasks (Sudhakar et al., 2019; Dinan et al., 2018; Urbanek et al., 2019). The language modeling aspects of generation like fluency and grammaticality are already learnt if pre-trained models are used.

These models are hard to fine-tune for sequence-to-sequence tasks such as machine translation, abstractive summarization etc. BART (Lewis et al., 2019) is a denoising autoencoder built with a sequence-to-sequence model and is particularly effective when fine tuned for text generation. Alternatively, T5 (Raffel et al., 2019) treats every NLP problem as a “text-to-text” problem, i.e. taking text as input and producing new text as output. Hence, it can be adapted to controllable text generation tasks. Dathathri et al. (2019) propose a Plug and Play Language Model (PPLM) for controllable language generation. It combines a pre-trained LM with one or more simple attribute classifiers that guide text generation without any further training of the LM. This is similar to the classifier feedback technique described in §2.6.3. Some of the other techniques described in this paper such as stochastic changes §2.2.2, external feedback §2.2.4 and §2.5.2, decompose §2.2.3 etc would be hard to incorporate into pre-trained language models without modifying the model architecture or fine-tuning entailing the significant cost of retraining.

## 2.5 Output

Here, we discuss the various techniques used to modulate the sequential output  $\mathbf{o}_t$  at each time step of the generator. This is marked as position 4 in Figure 2.1.

### 2.5.1 Attention

Attention is the most popular way of guiding the generation process. It is typically used to guide the generation process to focus on the source sequence (Bahdanau et al., 2015). The

attention calculating module takes as input the current hidden state  $\mathbf{h}_t$  of the generator at each time step  $t$ . The aim of this module is to determine a context vector  $\mathbf{c}_t$  that captures relevant source-side information to help predict the current target word  $\mathbf{y}_t$ . In case of *global attention*, all the hidden states of the encoder are considered to calculate the context vector  $\mathbf{c}_t$  (Luong et al., 2015a). This faces the the downside of expensive calculation especially for longer source sequences like documents. To overcome this challenge, *local attention* only chooses to focus only on a small subset of the source positions per target word. In this case,  $\mathbf{c}_t$  is calculated over a window of size  $D$  of the source hidden states.

Vaswani et al. (2017) view attention as a mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. This work proposes the simultaneous use of *scaled dot-product* attention which helps in parallelizing computation and a *multi-headed* attention which allows the model to jointly attend to information from different representation subspaces at different positions.

Sudhakar et al. (2019) use self-attention to control for style by simply adding the a special target style token in the source sequence. Dinan et al. (2018) also use transformers to attend over information from external document for guided dialogue response generation in their Two Stage model. (Zhang et al., 2018) uses the encoded representation of personas to compute the attention weights  $\mathbf{a}_t$  at a given time step of the decoder. The attention is reweighted according to the persona of the response to be generated in dialogue. So far, work has not been done to modulate the attention weights to control for attributes like style, topic, content etc.

### 2.5.2 External Feedback

The output latent space of the generator can be controlled by external feedback. Similar to changing the external input  $\mathbf{h}_0$ , the output latent space can also be changed using adversarial loss. In (Logeswaran et al., 2018a), an adversarial loss is used which encourages the generation realistic and attribute compatible sentences. The adversarial loss tries to match the distribution of sentence and attribute vector pairs  $(\mathbf{x}, \mathbf{s})$  where the sentence can either be a real or generated sentence. Similarly, in (Shen et al., 2017), a two discriminator losses in the style transfer task. Each discriminator is trained to distinguish between a sentence which came from the real target attribute distribution and a sentence that was transferred from source to target attribute. This work uses Professor-Forcing (Lamb et al., 2016) to match the hidden states of the generator and the discriminator. Gong et al. (2019) also control the output latent space by providing different types of rewards like style reward, semantic reward and fluency reward in the reinforcement learning setup. The discriminator used to obtain the adversarial loss has to be jointly trained with the generator.

### 2.5.3 Arithmetic or Linear Transform

Hoang et al. (2016) demonstrate three simple ways of changing the output  $\mathbf{o}_t$  of an RNN to control for meta information like topic, keywords etc. They show that you can add the control vector  $\mathbf{s}$  to  $\mathbf{o}_t$ . Hence the modified output  $\tilde{\mathbf{o}}_t$  is  $\tilde{\mathbf{o}}_t = \mathbf{o}_t + \mathbf{s}$ . Similarly, you can create  $\tilde{\mathbf{o}}_t$  by concatenating  $\mathbf{s}$  to  $\mathbf{o}_t$  ( $\tilde{\mathbf{o}}_t = [\mathbf{o}_t; \mathbf{s}]$ ). We can also build  $\tilde{\mathbf{o}}_t$  using a perceptron layer dependent on  $\mathbf{s}$  and  $\mathbf{o}_t$ . In this case,  $\tilde{\mathbf{o}}_t$  is given by  $\tilde{\mathbf{o}}_t = \tanh(\mathbf{W}_o \mathbf{o}_t + \mathbf{W}_s \mathbf{s} + \mathbf{b}_o)$ . In each of the three cases, the modified output  $\tilde{\mathbf{o}}_t$  is then projected to the vocabulary space to predict the token  $y_t$ .

## 2.6 Training Objective

In this section we describe various methods used to control the generation using objective functions. The output  $\mathbf{o}_t$  at each time step  $t$  of the generation process is projected to the vocabulary space using a linear transform ( $\tilde{\mathbf{o}}_t = \mathbf{W}_o \mathbf{o}_t + \mathbf{b}$ ). A token  $\hat{x}_t$  is predicted from the vocabulary by passing  $\mathbf{o}_t$  through a softmax function and taking the max value. The predicted token  $\hat{x}_t$  is compared with the reference token  $y_t$  using the loss function. This loss function can be tweaked to ensure that the generated text carries the desired control attributes.

### 2.6.1 General Loss Objectives

Here, we describe the loss objectives commonly used in natural language generation tasks. These loss objectives do not try to control for any attribute. Instead they try to ensure fluent, grammatical and diverse generations.

**Cross Entropy Loss** is the basic loss used to compare the generated tokens with the reference tokens and is used in all text generation process. At each time step  $t$ , the generation has to predict a token from the vocabulary. Hence, it could be seen as a classification problem with number of classes being equal to vocabulary size. The categorical cross entropy loss is given by:

$$-\sum_{c=1}^M y_{t,c} \log(p_{t,c})$$

where  $p_{t,c}$  is the probability of the token  $c$  at time step  $t$ . Note that  $p_t = \text{softmax}(\tilde{\mathbf{o}}_t)$  is the probability distribution over the vocabulary.

**Unlikelihood objective** maintains a set of negative candidates which is based on repeating tokens or n-grams and frequent tokens. This set is updated at each time step as tokens are generated. This works at both token and sequence level and the objective tries to minimize the repetitions in generations. This is used at train time in augmentation with the maximum likelihood objective and can be used for any task.

**Decoding Strategies:** are not used as a loss objective during training. Many of these objectives rely on post-hoc decoding strategies such as stochastic decoding which include Top  $k$ -sampling (Fan et al., 2018), nucleus sampling (Holtzman et al., 2020), or beam search variants (Paulus et al., 2018; Kulikov et al., 2019; Vijayakumar et al., 2018; Holtzman et al., 2018).

Specifically, we discuss the Diversity-Promoting objective which is used to generate a varied set of sentences given similar inputs. Particularly, Li et al. (2016a) use Maximum Mutual Information (MMI) as an objective function for the dialogue response generation task. Most generation systems use maximum likelihood objective but this objective additionally tries to reduce the proportion of generic responses. It is given by:

$$\hat{\mathbf{T}} = \operatorname{argmax}_T \{ \log p(\mathbf{T}|\mathbf{S}) - \lambda \log p(\mathbf{T}) \}$$

where  $\hat{\mathbf{T}}$  is the generated target sequence,  $\mathbf{T}$  is the reference target sequence and  $\mathbf{S}$  is the source sequence. The second term controls the generation of the high frequency or the generic target sequences. Note that this objective is only used during the inference and the generators are trained using cross entropy loss. Zhang et al. (2018), also use a diversity encouraging objective for dialogue response generation. They train a discriminator to calculate similarity between the source  $\mathbf{S}$  and target  $\mathbf{T}$  ( $D_\psi(\mathbf{T}, \mathbf{S})$ ), as well as between the source  $\mathbf{S}$  and the generated target  $\hat{\mathbf{T}}$  ( $D_\psi(\hat{\mathbf{T}}, \mathbf{S})$ ). They finally try to minimize the difference between  $D_\psi(\mathbf{T}, \mathbf{S})$  and  $D_\psi(\hat{\mathbf{T}}, \mathbf{S})$ .

### 2.6.2 KL Divergence

The Kullback-Leibler (KL) Divergence score, quantifies how much one probability distribution differs from another probability distribution. The KL divergence between two distributions  $\mathcal{Q}$  and  $\mathcal{P}$  is often stated using the following notation:

$$\text{KL}(\mathcal{P} \parallel \mathcal{Q})$$

where the operator “ $\parallel$ ” indicates *divergence* or  $\mathcal{P}$ ’s divergence from  $\mathcal{Q}$ . Note that KL Divergence is not symmetric i.e  $\text{KL}(\mathcal{P} \parallel \mathcal{Q}) \neq \text{KL}(\mathcal{Q} \parallel \mathcal{P})$ . KL divergence can be used to minimize the information loss while approximating a distribution. In text generation, the KL Divergence is combined with the evidence lower bound (ELBO) to approximately maximize the marginal likelihood of data  $p(\mathbf{x})$  which helps in better generations. This objective is used in variational autoencoders and its variants in combination with sampling techniques described in §2.2. This objective fits in the controllable text generation paradigm because it allows you to approximate the posterior distribution of the control variables in the latent  $\mathbf{z}$ -space.

### 2.6.3 Classifier Loss

This loss is specifically used to ensure that the generated tokens  $\hat{\mathbf{x}}$  comply with the control attributes  $\mathbf{s}$ . Note the difference between this loss and the external feedback loss used for the

*external input* module and the *output* module is that this loss operates at the token level and the external feedback loss works on the latent hidden representations.

In case of style transfer task, this loss is used to guide the generation process to output the target style tokens. Some works (Prabhumoye et al., 2018; Sudhakar et al., 2019; Hu et al., 2017) use this loss to discriminate between all the styles in their task (one versus all fashion). This type of design will suffer from low accuracy scores when the number of styles increases. To counter this problem, this loss can be setup to calculate if the generated sentence  $\hat{x}$  belongs to style  $s_1$  or not and similarly to calculate another separate loss term for each style (Chandu et al., 2019b). This type of loss design encounters increasing number of loss terms depending on the number of styles. The third way to motivate this loss term is to discriminating between a sentence  $x$  from data which belongs to style  $s_1$  and a generated sentence  $\hat{x}$  which belongs to the same style  $s_1$  (Yang et al., 2018b). Again, you would need as many loss terms as the number of styles in this case. All of these works use cross entropy loss function to measure their losses.

Hu et al. (2020a) use a classifier based loss in the visual storytelling task. The classifier is a pre-trained language model (Devlin et al., 2019) used to measure the coherence between generated sentences of the story. Particularly, the classifier takes as input two sentences at a time  $\hat{x}_1$  and  $\hat{x}_2$  and outputs a binary label which indicates if  $\hat{x}_2$  follows  $\hat{x}_1$ . In this case, the control variable is coherence in stories which is used to guide the generator to produce consistent sentences.

#### 2.6.4 Task Specific Loss

Depending on the end task and the attribute to be controlled, you can design different loss objectives to ensure that generations abide by the target attributes.

**Strategy Loss:** Zhou et al. (2020) use a dialogue strategy based objective to generate responses for negotiation tasks. This task has ground truth strategies that lead to better negotiations. This loss captures the probability of a particular strategy occurring for the next utterance given the dialogue history. It guides the generator to align the responses with particular strategies.

**Coverage Loss:** Generating repeated words or phrases is a common problem for text generation systems, and this becomes especially pronounced for multi-sentence text generation task such as abstractive document summarization. See et al. (2017) introduce a *coverage loss* which penalizes repeatedly attending to the same locations of the source document.

**Structure loss:** Li et al. (2018b) introduce two new loss objectives *structural compression* and *structural coverage* based on sentence-level attention. These objectives are specially designed for the task of abstractive document summarization. *structural compression* is used to generate

a sentence by compressing several specific source sentences and *structural coverage* is used to cover more salient information of the original document. These objectives leverage document structure in document summarization, and explore the effectiveness of capturing structural properties of document summarization by regularization of the generative model to generate more informative and concise summaries.

## 2.7 Discussion

**Discrete space issues:** The classifier loss (§2.6.3) is used to determine if the generated tokens  $\hat{\mathbf{x}}$  are in accordance with the target control attribute  $\mathbf{s}$ . To calculate the loss, the generated tokens  $\hat{\mathbf{x}}$  are provided as input to the classifier. If the tokens in this case are generated using the *argmax* then this function is not differentiable. Hence, passing tokens effectively to the classifier is a challenge.

In (Yu et al., 2017), the REINFORCE (Williams, 1992) algorithm is used and rewards are calculated using Monte Carlo search sampling for the next tokens. This technique is known to be unstable due to the high variance of the sampled gradient during training (Shen et al., 2017). Kusner and Hernández-Lobato (2016) introduce the Gumbel-softmax distribution as a solution. It approximates the multinomial distribution parameterized in terms of the softmax distribution. Here the predicted token is:

$$\hat{\mathbf{x}}_t = \text{softmax}(1/\tau(\hat{\mathbf{o}}_t + \mathbf{g}_t)),$$

where  $\hat{\mathbf{o}}_t$  is described in (§2.6),  $\tau$  is temperature parameter and  $\mathbf{g}_t$  is sampled independently from the Gumbel distribution. Hu et al. (2017) use this technique without sampling from the Gumbel distribution but by only training the temperature parameter.

**Combined module architecture:** It is also possible to combine techniques from multiple modules to control the generation process. Some of the prior works have successfully combined various modules here. Hu et al. (2017) combine stochastic changes (§2.2.2), KL Divergence loss (§2.6.2) and a classifier loss (§2.6.3). It adopts a variational auto-encoder along with KL divergence loss objective and further adds a discriminator loss which signifies if the generated sentence belong to the target attribute. As mentioned earlier, Romanov et al. (2019) combine the decomposition of the external input (§2.2.3) with external feedback provided to the external input (§2.2.4). External feedback is used to ensure that the decomposed latent sub-spaces represent the desired target attributes. Hu et al. (2018) establishes formal connections between generative adversarial networks (related to §2.5.2 and §2.6.3) and variational auto-encoders (related to §2.2.2 and §2.6.2). Determining the best possible combination of modules through empirical evaluation remains an open challenge.

## 2.8 Conclusion

In this chapter, we propose a new schema to organize the prior work in controllable text generation. The schema contains five modules, each of which plays an important role in the generation process. We detail the various techniques used to modulate each of the five modules to perform controllable text generation. We also provide theoretical understanding and qualitative analysis of these techniques. This understanding paves way to new architectures based on combinations of these modules. The future work can focus on empirical comparison of these techniques to gain an insight into their usefulness and strength.



## Chapter 3

# Style Transfer

Style is used to communicate in an economical, strategic and believable way. For example simply stating ‘I’m angry’ is less convincing than shouting ‘Damn!’ (Eckert, 2019). Yet, defining a style is a non-trivial task. A descriptive approach to defining a style is of very little use in a theory of language production, since it never makes clear why and how each style is formed out of words; nor does it indicate any systematicity behind the classification of styles. Additionally, classifying all the possible styles of text is an impossible task: One can imagine text characteristics that fit almost any adjective! (Hovy, 1987). Eckert (2019) investigates the social indexicalities that can contribute to the emergence of a particular style. An important point made here is that style develops as a contrast to the existing indexicalities. An example of this in the non-linguistic domain is that if everyone wears black all the time then there is no existence of a style. Style will only exist when at least one person decides to wear a different color or form of clothing.

Kang and Hovy (2019) adopts Hovy (1987)’s functional approach to define style by its pragmatics aspects and rhetorical goals, or personal and group characteristics of participants. This work categorizes style along the two axes of social participation and content coupledness. It further identifies demographic attributes such as gender, age, education etc as personal styles, and formality, politeness as interpersonal. Sentiment, humor, romance on the other hand have been identified as heavily content coupled styles. For the purpose of the experiments, we assume a group of examples of text that belong to same label as one style. For example a set of sentences from a comedy show intended to incite humor are considered to belong to humorous style. Similarly, sentences written by George Orwell would be considered to be written in the Orwellian author style. We acknowledge that a piece of text could be a mix of multiple styles. For example, Orwellian work is both written in the author style as well as satirical.

Style transfer is the task of rephrasing the text to contain specific stylistic properties without changing the intent or affect within the context. There is a constant debate in the community on what is considered as preserving the semantic content in the case of style transfer. In our opinion the evaluation of meaning preservation should be done using the downstream

application for which the style transfer is to be used. For example, when writing a customer review for a product or restaurant, the overall sentiment of the review should remain the same while changing the demographic attributes or politeness of the review. If the review complains or appreciates the food/service then the generated sentence should maintain the same. In a lenient evaluation it might be ok to change the name of the food item in the review. But such a mistake will not be appreciated if the downstream task is ordering food from a restaurant or products from Amazon. When generating sentences for orders in different styles, the quantity and the food item/product name should remain the same in the output.

The most popular application of style transfer is to generate diverse responses for dialogue systems. You can control politeness, authority, persona etc of the dialogue responses. Style transfer can also be used to control the politeness of email requests. We have automatically labelled a huge dataset of 1.39 million sentences from Enron email corpus (Yeh and Harnly, 2006) for politeness (Madaan et al., 2020b) and show effective transfer of non-polite email requests to polite. Story generation is another interesting application of style transfer. You can use style transfer to generate the story with different emotional endings (Peng et al., 2018) or as we show in (Chandu et al., 2019b), you can generate stories in different persona types. The use of style transfer in machine translation task has recently caught attention (Niu et al., 2017; Niu and Carpuat, 2019).

**Challenges:** The main challenge in style transfer task is the lack of parallel data. Very few datasets exist with sentences which are aligned in all styles (Rao and Tetreault, 2018). This also makes it hard to evaluate the generated sentences for the style transfer task. The other challenges include not having good definitions of style, the datasets for style transfer may contain confounding variables on which the sentences might depend on, there are no good evaluation metrics to evaluate both style transfer accuracy and meaning preservation in style, for style transfer techniques it is hard to disentangle the meaning of a sentence from its style.

**Overview:** We first describe the various tasks of style transfer in §3.1. We propose two novel approaches namely—*back-translation for style transfer* and *tag and generate* in §3.2. The experiments and results are showcased in §3.3 and a literature survey is presented in §3.4. The gender transfer task (§3.1.1), political slant task (§3.1.2) and the back-translation methodology (§3.2.1) is done in collaboration with Yulia Tsvetkov, Ruslan Salakhutdinov and Alan W Black (Prabhunoye et al., 2018). The work on politeness transfer described in §3.1.4 and §3.2.2 is done in collaboration with Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov and Alan W Black (Madaan et al., 2020b).

### 3.1 Tasks and Datasets

Much work in computational social science has shown that people’s personal and demographic characteristics—either publicly observable (e.g., age, gender) or private (e.g., religion, political

affiliation)—are revealed in their linguistic choices (Nguyen et al., 2016). There are practical scenarios, however, when these attributes need to be modulated or obfuscated. For example, some users may wish to preserve their anonymity online, for personal security concerns (Jardine, 2016), or to reduce stereotype threat (Spencer et al., 1999). Modulating authors’ attributes while preserving meaning of sentences can also help generate demographically-balanced training data for a variety of downstream applications.

Moreover, prior work has shown that the quality of language identification and POS tagging degrades significantly on African American Vernacular English (Blodgett et al., 2016; Jørgensen et al., 2015); YouTube’s automatic captions have higher error rates for women and speakers from Scotland (Rudinger et al., 2017). Synthesizing balanced training data—using style transfer techniques—is a plausible way to alleviate bias present in existing NLP technologies.

We thus focus on two tasks that have practical and social-good applications, and also accurate style classifiers. To position our method with respect to prior work, we employ a third task of sentiment transfer, which was used in two state-of-the-art approaches to style transfer (Hu et al., 2017; Shen et al., 2017). We describe the four tasks and associated dataset statistics below. The methodology that we advocate is general and can be applied to other styles, for transferring various social categories, types of bias, and in multi-class settings.

### 3.1.1 Gender Transfer

In sociolinguistics, gender is known to be one of the most important social categories driving language choice (Eckert and McConnell-Ginet, 2003; Lakoff and Bucholtz, 2004; Coates, 2015; Tannen, 1991, 1993; Kendall et al., 1997; Eckert and McConnell-Ginet, 2003; Lakoff and Bucholtz, 2004; Coates, 2015). Numerous computational studies have also explored how gender is manifested in language of social media users (Rao et al., 2010; Burger et al., 2011; Peersman et al., 2011; Bergsma and Van Durme, 2013; Flekova and Gurevych, 2013; Bamman et al., 2014; Volkova et al., 2015; Carpenter et al., 2016, *inter alia*). Gender-induced style differences include, for example, that women are more likely to use pronouns, emotion words (like *sad*, *love*, and *glad*), interjections (*ah*, *hmmmm*, *ugh*), emoticons, and abbreviations associated with online discourse (*lol*, *omg*), while men tend to use higher frequency standard dictionary words, proper names (e.g., the names of sports teams), numbers, technology words, and links (Bamman et al., 2014). Reddy and Knight (2016) proposed a heuristic-based method to obfuscate gender of a writer. This method uses statistical association measures to identify gender-salient words and substitute them with synonyms typically of the opposite gender. This simple approach produces highly fluent, meaning-preserving sentences, but does not allow for more general rephrasing of sentence beyond single-word substitutions. In our work, we adopt this task of transferring the author’s gender and adapt it to our experimental settings.

We used Reddy and Knight’s (2016) dataset of reviews from Yelp annotated for two genders corresponding to markers of sex.<sup>1</sup> We split the reviews to sentences, preserving the original gender labels. To keep only sentences that are strongly indicative of a gender, we then filtered out gender-neutral sentences (e.g., *thank you*) and sentences whose likelihood to be written by authors of one gender is lower than 0.7.<sup>2</sup>

### 3.1.2 Political Slant Transfer

Our second dataset is comprised of top-level comments on Facebook posts from all 412 current members of the United States Senate and House who have public Facebook pages (Voigt et al., 2018).<sup>3</sup> Only top-level comments that directly respond to the post are included. Every comment to a Congressperson is labeled with the Congressperson’s party affiliation: democratic or republican. Topic and sentiment in these comments reveal commenter’s political slant. For example, *defund them all, especially when it comes to the illegal immigrants* . and *thank u james, praying for all the work u do* . are republican, whereas *on behalf of the hard-working nh public school teachers- thank you !* and *we need more strong voices like yours fighting for gun control* . represent examples of democratic sentences. Our task is to preserve intent of the commenter (e.g., to thank their representative), but to modify their observable political affiliation, as in the example in Figure 3.3. We preprocessed and filtered the comments similarly to the gender-annotated corpus above.

### 3.1.3 Sentiment Modification

To compare our work with the state-of-the-art approaches of style transfer for non-parallel corpus we perform sentiment transfer, replicating the models and experimental setups of Hu et al. (2017) and Shen et al. (2017). Given a positive Yelp review, a style transfer model will generate a similar review but with an opposite sentiment. We used Shen et al.’s (2017) corpus of reviews from Yelp. They have followed the standard practice of labeling the reviews with rating of higher than three as positive and less than three as negative. They have also split the reviews to sentences and assumed that the sentence has the same sentiment as the review.

**Dataset statistics:** We summarize below corpora statistics for the three tasks: transferring gender, political slant, and sentiment. The dataset for sentiment modification task was used as described in (Shen et al., 2017). We split Yelp and Facebook corpora into four disjoint parts each: (1) a training corpus for training a style classifier (*class*); (2) a training corpus (*train*) used for training the style-specific generative model described in §3.2.1; (3) development and

<sup>1</sup>We note that gender may be considered along a spectrum (Eckert and McConnell-Ginet, 2003), but use gender as a binary variable due to the absence of corpora with continuous-valued gender annotations.

<sup>2</sup>We did not experiment with other threshold values.

<sup>3</sup>The posts and comments are all public; however, to protect the identity of Facebook users in this dataset Voigt et al. (2018) have removed all identifying user information as well as Facebook-internal information such as User IDs and Post IDs, replacing these with randomized ID numbers.

Style	<i>class</i>	<i>train</i>	<i>dev</i>	<i>test</i>
gender	2.57M	2.67M	4.5K	535K
political	80K	540K	4K	56K
sentiment	2M	444K	63.5K	127K

TABLE 3.1: Sentence count in style-specific corpora.

(4) test sets. We have removed from training corpora *class* and *train* all sentences that overlap with development and test corpora. Corpora sizes are shown in Table 3.1. Table 3.2 shows the approximate vocabulary sizes used for each dataset. The vocabulary is the same for both the styles in each experiment. Table 3.3 summarizes sentence statistics. All the sentences have maximum length of 50 tokens.

Style	gender	political	sentiment
Vocabulary	20K	20K	10K

TABLE 3.2: Vocabulary sizes of the datasets.

Style	Avg. Length	%data
male	18.08	50.00
female	18.21	50.00
republican	16.18	50.00
democratic	16.01	50.00
negative	9.66	39.81
positive	8.45	60.19

TABLE 3.3: Average sentence length and class distribution of style corpora.

### 3.1.4 Politeness Transfer

For the politeness transfer task, we focus on sentences in which the speaker communicates a requirement that the listener needs to fulfill. Common examples include imperatives “*Let’s stay in touch*” and questions that express a proposal “*Can you call me when you get back?*”. Following Jurafsky et al. (1997), we use the umbrella term “action-directives” for such sentences. The goal of this task is to convert action-directives to polite requests. While there can be more than one way of making a sentence polite, for the above examples, adding gratitude (“*Thanks and let’s stay in touch*”) or counterfactuals (“*Could you please call me when you get back?*”) would make them polite (Danescu-Niculescu-Mizil et al., 2013).

**Data Preparation** The Enron corpus (Klimt and Yang, 2004) consists of a large set of email conversations exchanged by the employees of the Enron corporation. Emails serve as a medium for exchange of requests, serving as an ideal application for politeness transfer. We begin by

pre-processing the raw Enron corpus following Shetty and Adibi (2004). The first set of pre-processing steps and de-duplication yielded a corpus of roughly 2.5 million sentences.<sup>4</sup> Further pruning<sup>5</sup> led to a cleaned corpus of over 1.39 million sentences. Finally, we use a politeness classifier (Niu and Bansal, 2018b) to assign politeness scores to these sentences and filter them into ten buckets based on the score ( $P_0$ - $P_9$ ; Figure 3.1). All the buckets are further divided into train, test, and dev splits (in a 80:10:10 ratio).

For our experiments, we assumed all the sentences with a politeness score of over 90% by the classifier to be polite, also referred as the  $P_9$  bucket (marked in green in Figure 3.1). We use the train-split of the  $P_9$  bucket of over 270K polite sentences as the training data for the politeness transfer task. Since the goal of the task is making action directives more polite, we manually curate a test set comprising of such sentences from test splits across the buckets. We first train a classifier on the switchboard corpus (Jurafsky et al., 1997) to get dialog state tags and filter sentences that have been labeled as either action-directive or quotation.<sup>6</sup> Further, we use human annotators to manually select the test sentences. The annotators had a Fleiss’s Kappa score ( $\kappa$ ) of 0.77<sup>7</sup> and curated a final test set of 800 sentences.

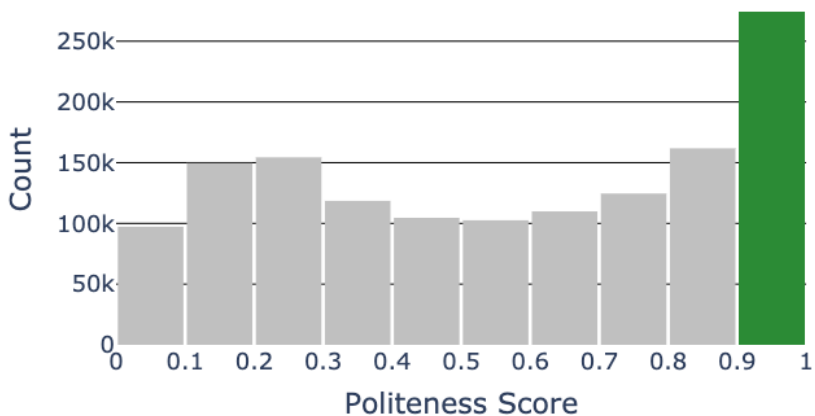


FIGURE 3.1: Distribution of Politeness Scores for the Enron Corpus

In Figure 3.2, we examine the two extreme buckets with politeness scores of  $< 10\%$  ( $P_0$  bucket) and  $> 90\%$  ( $P_9$  bucket) from our corpus by plotting 10 of the top 30 words occurring in each bucket. We clearly notice that words in the  $P_9$  bucket are closely linked to polite style, while words in the  $P_0$  bucket are mostly content words. This substantiates our claim that the task of politeness transfer is fundamentally different from other attribute transfer tasks like sentiment where both the polarities are clearly defined.

<sup>4</sup>Pre-processing also involved steps for tokenization (done using spacy (Honnibal and Montani, 2017)) and conversion to lower case.

<sup>5</sup>We prune the corpus by removing the sentences that 1) were less than 3 words long, 2) had more than 80% numerical tokens, 3) contained email addresses, or 4) had repeated occurrences of spurious characters.

<sup>6</sup>We used AWD-LSTM based classifier for classification of action-directive.

<sup>7</sup>The score was calculated for 3 annotators on a sample set of 50 sentences.

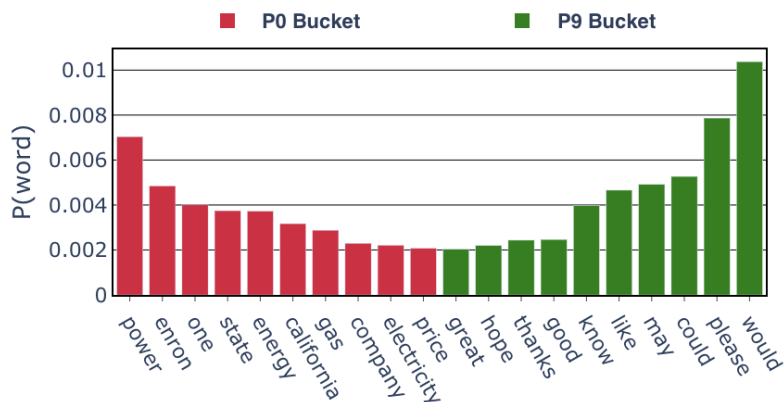


FIGURE 3.2: Probability of occurrence for 10 of the most common 30 words in the  $P_0$  and  $P_9$  data buckets

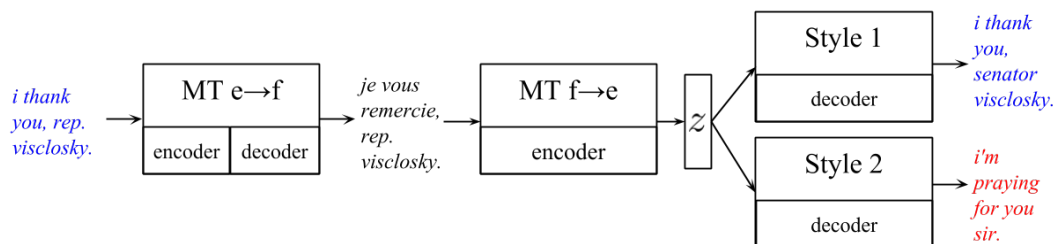


FIGURE 3.3: Style transfer pipeline: to rephrase a sentence and reduce its stylistic characteristics, the sentence is back-translated. Then, separate style-specific generators are used for style transfer.

## 3.2 Methodology

### 3.2.1 Back-translation

We introduce a novel approach to transferring style of a sentence while better preserving its meaning. We hypothesize—relying on the study of [Rabinovich et al. \(2017\)](#) who showed that author characteristics are significantly obfuscated by both manual and automatic machine translation—that grounding in back-translation is a plausible approach to rephrase a sentence while reducing its stylistic properties. We thus first use back-translation to rephrase the sentence and reduce the effect of the original style; then, we generate from the latent representation, using separate style-specific generators controlling for style.

Given two datasets  $\mathbf{X}_1 = \{\mathbf{x}_1^{(1)}, \dots, \mathbf{x}_1^{(n)}\}$  and  $\mathbf{X}_2 = \{\mathbf{x}_2^{(1)}, \dots, \mathbf{x}_2^{(n)}\}$  which represent two different styles  $s_1$  and  $s_2$ , respectively, our task is to generate sentences of the desired style while preserving the meaning of the input sentence. Specifically, we generate samples of dataset  $\mathbf{X}_1$  such that they belong to style  $s_2$  and samples of  $\mathbf{X}_2$  such that they belong to style  $s_1$ . We denote the output of dataset  $\mathbf{X}_1$  transferred to style  $s_2$  as  $\hat{\mathbf{X}}_1 = \{\hat{\mathbf{x}}_1^{(1)}, \dots, \hat{\mathbf{x}}_1^{(n)}\}$  and the output of dataset  $\mathbf{X}_2$  transferred to style  $s_1$  as  $\hat{\mathbf{X}}_2 = \{\hat{\mathbf{x}}_2^{(1)}, \dots, \hat{\mathbf{x}}_2^{(n)}\}$ .

Hu et al. (2017) and Shen et al. (2017) introduced state-of-the-art style transfer models that use variational auto-encoders (Kingma and Welling, 2014, VAEs) and cross-aligned auto-encoders, respectively, to model a latent content variable  $z$ . The latent content variable  $z$  is a code which is not observed. The generative model conditions on this code during the generation process. Our aim is to design a latent code  $z$  which (1) represents the meaning of the input sentence grounded in back-translation and (2) weakens the style attributes of author’s traits. To model the former, we use neural machine translation. Prior work has shown that the process of translating a sentence from a source language to a target language retains the meaning of the sentence but does not preserve the stylistic features related to the author’s traits (Rabinovich et al., 2017). We hypothesize that a latent code  $z$  obtained through back-translation will normalize the sentence and devoid it from style attributes specific to author’s traits.

Figure 3.3 shows the overview of the proposed method. In our framework, we first train a machine translation model from source language  $e$  to a target language  $f$ . We also train a back-translation model from  $f$  to  $e$ . Let us assume the styles  $s_1$  and  $s_2$  correspond to DEMOCRATIC and REPUBLICAN style, respectively. In Figure 3.3, the input sentence *i thank you, rep. visclosky.* is labeled as DEMOCRATIC. We translate the sentence using the  $e \rightarrow f$  machine translation model and generate the parallel sentence in the target language  $f$ : *je vous remercie, rep. visclosky.* Using the fixed encoder of the  $f \rightarrow e$  machine translation model, we encode this sentence in language  $f$ . The hidden representation created by this encoder of the back-translation model is used as  $z$ . We condition our generative models on this  $z$ . We then train two separate decoders for each style  $s_1$  and  $s_2$  to generate samples in these respective styles in source language  $e$ . Hence the sentence could be translated to the REPUBLICAN style using the decoder for  $s_2$ . For example, the sentence *i’m praying for you sir.* is the REPUBLICAN version of the input sentence and *i thank you, senator visclosky.* is the more DEMOCRATIC version of it.

Note that in this setting, the machine translation and the encoder of the back-translation model remain fixed. They are not dependent on the data we use across different tasks. This facilitates re-usability and spares the need of learning separate models to generate  $z$  for a new style data.

The Back-translation technique modifies the training objective module (§2.6) of the schema described in ch. 2. Specifically, it uses an additional classifier loss objective to guide the generator towards the target style.

## Meaning-Grounded Representation

In this section we describe how we learn the latent content variable  $z$  using back-translation. The  $e \rightarrow f$  machine translation and  $f \rightarrow e$  back-translation models are trained using a sequence-to-sequence framework (Sutskever et al., 2014; Bahdanau et al., 2015) with style-agnostic corpus. The style-specific sentence *i thank you, rep. visclosky.* in source language  $e$  is translated to the target language  $f$  to get *je vous remercie, rep. visclosky.* The individual tokens of this sentence are then encoded using the encoder of the  $f \rightarrow e$  back-translation model. The learned hidden representation is  $z$ .



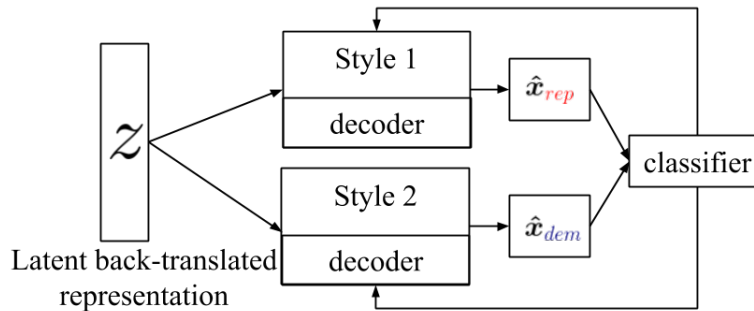


FIGURE 3.4: The latent representation from back-translation and the style classifier feedback are used to guide the style-specific generators.

Formally, let  $\theta_E$  represent the parameters of the encoder of  $f \rightarrow e$  translation system. Then  $z$  is given by:

$$z = \text{Encoder}(x_f; \theta_E) \quad (3.1)$$

where,  $x_f$  is the sentence  $x$  in language  $f$ . Specifically,  $x_f$  is the output of  $e \rightarrow f$  translation system when  $x_e$  is given as input. Since  $z$  is derived from a non-style specific process, this Encoder is not style specific.

### Style-Specific Generation

Figure 3.4 shows the architecture of the generative model for generating different styles. Using the encoder embedding  $z$ , we train multiple decoders for each style. The sentence generated by a decoder is passed through the classifier. The loss of the classifier for the generated sentence is used as feedback to guide the decoder for the generation process. The target attribute of the classifier is determined by the decoder from which the output is generated. For example, in the case of DEMOCRATIC decoder, the target attribute is DEMOCRATIC and for the REPUBLICAN decoder the target is REPUBLICAN.

### Style Classifiers

We train a convolutional neural network (CNN) classifier to accurately predict the given style. We also use it to evaluate the error in the generated samples for the desired style. We train the classifier in a supervised manner. The classifier accepts either discrete or continuous tokens as inputs. This is done such that the generator output can be used as input to the classifier. We need labeled examples to train the classifier such that each instance in the dataset  $\mathbf{X}$  should have a label in the set  $s = \{s_1, s_2\}$ . Let  $\theta_C$  denote the parameters of the classifier. The objective to train the classifier is given by:

$$\mathcal{L}_{class}(\theta_C) = \mathbb{E}_{\mathbf{X}}[\log q_C(s|\mathbf{x})]. \quad (3.2)$$

To improve the accuracy of the classifier, we augment the classifier’s inputs with style-specific lexicons. We concatenate binary style indicators to each input word embedding in the classifier. The indicators are set to 1 if the input word is present in a style-specific lexicon; otherwise they are set to 0. Style lexicons are extracted using the log-odds ratio informative Dirichlet prior (Monroe et al., 2008), a method that identifies words that are statistically overrepresented in each of the categories.

### Generator Learning

We use a bidirectional LSTM to build the decoders which generate the sequence of tokens  $\hat{\mathbf{x}} = \{x_1, \dots, x_T\}$ . The sequence  $\hat{\mathbf{x}}$  is conditioned on the latent code  $\mathbf{z}$  (in our case, on the machine translation model). In this work we use a corpus translated to French by the machine translation system as the input to the encoder of the back-translation model. The same encoder is used to encode sentences of both styles. The representation created by this encoder is given by Eq. 3.1. Samples are generated as follows:

$$\hat{\mathbf{x}} \sim \mathbf{z} = p(\hat{\mathbf{x}}|\mathbf{z}) \quad (3.3)$$

$$= \prod_t p(\hat{x}_t | \hat{\mathbf{x}}^{<t}, \mathbf{z}) \quad (3.4)$$

where,  $\hat{\mathbf{x}}^{<t}$  are the tokens generated before  $\hat{x}_t$ .

Tokens are discrete and non-differentiable. This makes it difficult to use a classifier, as the generation process samples discrete tokens from the multinomial distribution parametrized using softmax function at each time step  $t$ . This non-differentiability, in turn, breaks down gradient propagation from the discriminators to the generator. Instead, following Hu et al. (2017) we use a continuous approximation based on softmax, along with the temperature parameter which anneals the softmax to the discrete case as training proceeds. To create a continuous representation of the output of the generative model which will be given as an input to the classifier, we use:

$$\hat{x}_t \sim \text{softmax}(\mathbf{o}_t/\tau),$$

where,  $\mathbf{o}_t$  is the output of the generator and  $\tau$  is the temperature which decreases as the training proceeds. Let  $\theta_G$  denote the parameters of the generators. Then the reconstruction loss is calculated using the cross entropy function, given by:

$$\mathcal{L}_{recon}(\theta_G; \mathbf{x}) = \mathbb{E}_{q_E(\mathbf{z}|\mathbf{x})}[\log p_{gen}(\mathbf{x}|\mathbf{z})] \quad (3.5)$$

Here, the back-translation encoder  $E$  creates the latent code  $\mathbf{z}$  by:

$$\mathbf{z} = E(\mathbf{x}) = q_E(\mathbf{z}|\mathbf{x}) \quad (3.6)$$



FIGURE 3.5: Our proposed approach: *tag* and *generate*. The tagger infers the interpretable style free sentence  $z(\mathbf{x}_i)$  for an input  $\mathbf{x}_i^{(1)}$  in source style  $\mathcal{S}_1$ . The generator transforms  $\mathbf{x}_i^{(1)}$  into  $\hat{\mathbf{x}}_i^{(2)}$  which is in target style  $\mathcal{S}_2$ .

The generative loss  $\mathcal{L}_{gen}$  is then given by:

$$\min_{\theta_{gen}} \mathcal{L}_{gen} = \mathcal{L}_{recon} + \lambda_c \mathcal{L}_{class} \quad (3.7)$$

where  $\mathcal{L}_{recon}$  is given by Eq. 3.5,  $\mathcal{L}_{class}$  is given by Eq. 3.2 and  $\lambda_c$  is a balancing parameter.

We also use global attention of (Luong et al., 2015b) to aid our generators. At each time step  $t$  of the generation process, we infer a variable length alignment vector  $\mathbf{a}_t$ :

$$\mathbf{a}_t = \frac{\exp(\text{score}(\mathbf{h}_t, \bar{\mathbf{h}}_s))}{\sum_{s'} \exp(\text{score}(\mathbf{h}_t, \bar{\mathbf{h}}_{s'}))} \quad (3.8)$$

$$\text{score}(\mathbf{h}_t, \bar{\mathbf{h}}_s) = \text{dot}(\mathbf{h}_t^T, \bar{\mathbf{h}}_s), \quad (3.9)$$

where  $\mathbf{h}_t$  is the current target state and  $\bar{\mathbf{h}}_s$  are all source states. While generating sentences, we use the attention vector to replace unknown characters (UNK) using the copy mechanism in (See et al., 2017).

### 3.2.2 Tag and Generate

We are given non-parallel samples of sentences  $\mathbf{X}_1 = \{\mathbf{x}_1^{(1)} \dots \mathbf{x}_n^{(1)}\}$  and  $\mathbf{X}_2 = \{\mathbf{x}_1^{(2)} \dots \mathbf{x}_m^{(2)}\}$  from styles  $\mathcal{S}_1$  and  $\mathcal{S}_2$  respectively. The objective of the task is to efficiently generate samples  $\hat{\mathbf{X}}_1 = \{\hat{\mathbf{x}}_1^{(2)} \dots \hat{\mathbf{x}}_n^{(2)}\}$  in the target style  $\mathcal{S}_2$ , conditioned on samples in  $\mathbf{X}_1$ . For a style  $\mathcal{S}_v$  where  $v \in \{1, 2\}$ , we begin by learning a set of phrases ( $\Gamma_v$ ) which characterize the style  $\mathcal{S}_v$ . The presence of phrases from  $\Gamma_v$  in a sentence  $\mathbf{x}_i$  would associate the sentence with the style  $\mathcal{S}_v$ . For example, phrases like “pretty good” and “worth every penny” are characteristic of the “positive” style in the case of sentiment transfer task.

We propose a two staged approach where we first infer a sentence  $z(\mathbf{x}_i)$  from  $\mathbf{x}_i^{(1)}$  using a model, the tagger. The goal of the tagger is to ensure that the sentence  $z(\mathbf{x}_i)$  is agnostic to the original style ( $\mathcal{S}_1$ ) of the input sentence. Conditioned on  $z(\mathbf{x}_i)$ , we then generate the transferred sentence  $\hat{\mathbf{x}}_i^{(2)}$  in the target style  $\mathcal{S}_2$  using another model, the generator. The intermediate variable  $z(\mathbf{x}_i)$  is also seen in other style-transfer methods. Shen et al. (2017); Prabhumoye et al. (2018); Yang et al. (2018b); Hu et al. (2017) transform the input  $\mathbf{x}_i^{(v)}$  to a latent representation  $z(\mathbf{x}_i)$  which (ideally) encodes the content present in  $\mathbf{x}_i^{(v)}$  while being agnostic to style  $\mathcal{S}_v$ . In these cases  $z(\mathbf{x}_i)$  encodes the input sentence in a continuous latent space whereas for us  $z(\mathbf{x}_i)$

manifests in the surface form. The ability of our pipeline to generate observable intermediate outputs  $z(\mathbf{x}_i)$  makes it somewhat more interpretable than those other methods.

We train two independent systems for the tagger & generator which have complimentary objectives. The former identifies the style attribute markers  $a(x_i^{(1)})$  from source style  $\mathcal{S}_1$  and either replaces them with a positional token called [TAG] or merely adds these positional tokens without removing any phrase from the input  $x_i^{(1)}$ . This particular capability of the model enables us to generate these tags in an input that is devoid of any attribute marker (i.e.  $a(x_i^{(1)}) = \{\}$ ). This is one of the major differences from prior works which mainly focus on removing source style attributes and then replacing them with the target style attributes. It is especially critical for tasks like politeness transfer where the transfer takes place from a non-polite sentence. This is because in such cases we may need to add new phrases to the sentence rather than simply replace existing ones. The generator is trained to generate sentences  $\hat{\mathbf{x}}_i^{(2)}$  in the target style by replacing these [TAG] tokens with stylistically relevant words inferred from target style  $\mathcal{S}_2$ . Even though we have non-parallel corpora, both systems are trained in a supervised fashion as sequence-to-sequence models with their own distinct pairs of inputs & outputs. To create parallel training data, we first estimate the style markers  $\Gamma_v$  for a given style  $\mathcal{S}_v$  & then use these to curate style free sentences with [TAG] tokens.

Figure 3.5 shows the overall pipeline of the proposed approach. In the first example  $\mathbf{x}_1^{(1)}$ , where there is no clear style attribute present, our model adds the [TAG] token in  $z(\mathbf{x}_1)$ , indicating that a target style marker should be generated in this position. On the contrary, in the second example, the terms “ok” and “bland” are markers of negative sentiment and hence the tagger has replaced them with [TAG] tokens in  $z(\mathbf{x}_2)$ . We can also see that the inferred sentence in both the cases is free of the original and target styles. The structural bias induced by this two staged approach is helpful in realizing an interpretable style free tagged sentence that explicitly encodes the content. In the following sections we discuss in detail the methodologies involved in (1) estimating the relevant attribute markers for a given style, (2) tagger, and (3) generator modules of our approach.

## Estimating Style Phrases

Drawing from Li et al. (2018a), we propose a simple approach based on n-gram tf-idfs to estimate the set  $\Gamma_v$ , which represents the style markers for style  $v$ . For a given corpus pair  $\mathbf{X}_1, \mathbf{X}_2$  in styles  $\mathcal{S}_1, \mathcal{S}_2$  respectively we first compute a probability distribution  $p_1^2(w)$  over the n-grams  $w$  present in both the corpora (Eq. 3.11). Intuitively,  $p_1^2(w)$  is proportional to the probability of sampling an n-gram present in both  $\mathbf{X}_1, \mathbf{X}_2$  but having a much higher tf-idf value in  $\mathbf{X}_2$  relative to  $\mathbf{X}_1$ . This is how we define the impactful style markers for style  $\mathcal{S}_2$ .

$$\eta_1^2(w) = \frac{\frac{1}{m} \sum_{i=1}^m \text{tf-idf}(w, \mathbf{x}_i^{(2)})}{\frac{1}{n} \sum_{j=1}^n \text{tf-idf}(w, \mathbf{x}_j^{(1)})} \quad (3.10)$$

$$p_1^2(w) = \frac{\eta_1^2(w)^\gamma}{\sum_{w'} \eta_1^2(w')^\gamma} \quad (3.11)$$

where,  $\eta_1^2(w)$  is the ratio of the mean tf-idfs for a given n-gram  $w$  present in both  $\mathbf{X}_1, \mathbf{X}_2$  with  $|\mathbf{X}_1| = n$  and  $|\mathbf{X}_2| = m$ . Words with higher values for  $\eta_1^2(w)$  have a higher mean tf-idf in  $\mathbf{X}_2$  vs  $\mathbf{X}_1$ , and thus are more characteristic of  $\mathcal{S}_2$ . We further smooth and normalize  $\eta_1^2(w)$  to get  $p_1^2(w)$ . Finally, we estimate  $\Gamma_2$  by

$$\Gamma_2 = \{w : p_1^2(w) \geq k\}$$

In other words,  $\Gamma_2$  consists of the set of phrases in  $\mathbf{X}_2$  above a given style impact  $k$ .  $\Gamma_1$  is computed similarly where we use  $p_2^1(w), \eta_2^1(w)$ .

### Style Invariant Tagged Sentence

The tagger model (with parameters  $\theta_t$ ) takes as input the sentences in  $\mathbf{X}_1$  and outputs  $\{z(\mathbf{x}_i) : \mathbf{x}_i^{(1)} \in \mathbf{X}_1\}$ . Depending on the style transfer task, the tagger is trained to either (1) identify and replace style attributes  $a(\mathbf{x}_i^{(1)})$  with the token tag [TAG] (replace-tagger) or (2) add the [TAG] token at specific locations in  $\mathbf{x}_i^{(1)}$  (add-tagger). In both the cases, the [TAG] tokens indicate positions where the generator can insert phrases from the target style  $\mathcal{S}_2$ . Finally, we use the distribution  $p_1^2(w)/p_2^1(w)$  over  $\Gamma_2/\Gamma_1$  (§3.2.2) to draw samples of attribute-markers that would be replaced with the [TAG] token during the creation of training data.

The first variant, replace-tagger, is suited for a task like sentiment transfer where almost every sentence has some attribute markers  $a(\mathbf{x}_i^{(1)})$  present in it. In this case the training data comprises of pairs where the input is  $\mathbf{X}_1$  and the output is  $\{z(\mathbf{x}_i) : \mathbf{x}_i^{(1)} \in \mathbf{X}_1\}$ . The loss objective for replace-tagger is given by  $\mathcal{L}_r(\theta_t)$  in Eq. 3.12.

$$\mathcal{L}_r(\theta_t) = - \sum_{i=1}^{|\mathbf{X}_1|} \log P_{\theta_t}(z(\mathbf{x}_i) | \mathbf{x}_i^{(1)}; \theta_t) \quad (3.12)$$

The second variant, add-tagger, is designed for cases where the transfer needs to happen from style neutral sentences to the target style. That is,  $\mathbf{X}_1$  consists of style neutral sentences whereas  $\mathbf{X}_2$  consists of sentences in the target style. Examples of such a task include the tasks of politeness transfer (introduced in this paper) and caption style transfer (used by Li et al. (2018a)). In such cases, since the source sentences have no attribute markers to remove, the tagger learns to add [TAG] tokens at specific locations suitable for emanating style words in the target style.

The training data (Figure 3.6) for the add-tagger is given by pairs where the input is  $\{\mathbf{x}_i^{(2)} \setminus a(\mathbf{x}_i^{(2)}) : \mathbf{x}_i^{(2)} \in \mathbf{X}_2\}$  and the output is  $\{z(\mathbf{x}_i) : \mathbf{x}_i^{(2)} \in \mathbf{X}_2\}$ . Essentially, for the input we take samples

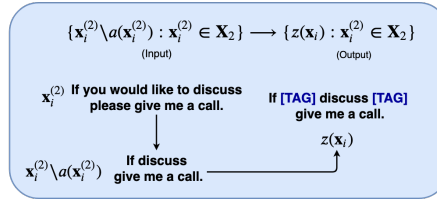


FIGURE 3.6: Creation of training data for add-tagger.

$x_i^{(2)}$  in the target style  $\mathcal{S}_2$  and explicitly remove style phrases  $a(x_i^{(2)})$  from it. For the output we replace the same phrases  $a(x_i^{(2)})$  with [TAG] tokens. As indicated in Figure 3.6, we remove the style phrases “you would like to” and “please” and replace them with [TAG] in the output. Note that we only use samples from  $\mathbf{X}_2$  for training the add-tagger; samples from the style neutral  $\mathbf{X}_1$  are not involved in the training process at all. For example, in the case of politeness transfer, we only use the sentences labeled as “polite” for training. In effect, by training in this fashion, the tagger learns to add [TAG] tokens at appropriate locations in a style neutral sentence. The loss objective ( $\mathcal{L}_a$ ) given by Eq. 3.13 is crucial for tasks like politeness transfer where one of the styles is poorly defined.

$$\mathcal{L}_a(\theta_t) = - \sum_{i=1}^{|\mathbf{X}_1|} \log P_{\theta_t}(z(\mathbf{x}_i) | \mathbf{x}_i^{(2)} \setminus a(\mathbf{x}_i^{(2)}); \theta_t) \quad (3.13)$$

### Style Targeted Generation

The training for the generator model is complimentary to that of the tagger, in the sense that the generator takes as input the tagged output  $z(\mathbf{x}_i)$  inferred from the source style and modifies the [TAG] tokens to generate the desired sentence  $\hat{\mathbf{x}}_i^{(v)}$  in the target style  $\mathcal{S}_v$ .

$$\mathcal{L}(\theta_g) = - \sum_{i=1}^{|\mathbf{X}_v|} \log P_{\theta_g}(\mathbf{x}_i^{(v)} | z(\mathbf{x}_i); \theta_g) \quad (3.14)$$

The training data for transfer into style  $\mathcal{S}_v$  comprises of pairs where the input is given by  $\{z(\mathbf{x}_i) : x_i^{(v)} \in \mathbf{X}_v, v \in \{1, 2\}\}$  and the output is  $\mathbf{X}_v$ , i.e. it is trained to transform a style agnostic representation into a style targeted sentence. Since the generator has no notion of the original style and it is only concerned with the style agnostic representation  $z(\mathbf{x}_i)$ , it is convenient to disentangle the training for tagger & generator.

Finally, we note that the location at which the tags are generated has a significant impact on the distribution over style attributes (in  $\Gamma_2$ ) that are used to fill the [TAG] token at a particular position. Hence, instead of using a single [TAG] token, we use a set of positional tokens  $[\text{TAG}]_t$  where  $t \in \{0, 1, \dots, T\}$  for a sentence of length  $T$ . By training both tagger and generator with

Experiment	CAE	BST
Gender	<b>60.40</b>	57.04
Political slant	75.82	<b>88.01</b>
Sentiment	80.43	<b>87.22</b>

TABLE 3.4: Accuracy of the style transfer in sentences generated by the BST and CAE models.

these positional  $[\text{TAG}]_t$  tokens we enable them to easily realize different distributions of style attributes for different positions in a sentence. For example, in the case of politeness transfer, the tags added at the beginning ( $t = 0$ ) will almost always be used to generate a token like “Would it be possible ...” whereas for a higher  $t$ ,  $[\text{TAG}]_t$  may be replaced with a token like “thanks” or “sorry.”

### 3.3 Experiments

**Translation quality:** The BLEU scores achieved for English–French MT system is 32.52 and for French–English MT system is 31.11; these are strong translation systems. We deliberately chose a European language close to English for which massive amounts of parallel data are available and translation quality is high, to concentrate on the style generation, rather than improving a translation system.<sup>8</sup>

Evaluating style transfer techniques is hard. We have to not only evaluate the generations for the success of style transfer but also if the generated sentence maintains the same meaning as the input sentence. Additionally, we must also evaluate if the generations are syntactically and grammatically sound. Both automatic evaluation and human judgments are used to evaluate style transfer systems along the three dimensions of: (1) Style transfer accuracy, measuring the proportion of our models’ outputs that generate sentences of the desired style. (2) Preservation of meaning. (3) Fluency, measuring the readability and the naturalness of the generated sentences.

We denote the Cross-aligned Auto-Encoder model (Shen et al., 2017) as CAE, the **Delete, Retrieve, Generate** model (Li et al., 2018a) as DRG, our back-translation model as **Back-translation for Style Transfer** (BST) and our **Tag and Generate** approach as TaG.

#### 3.3.1 Style Transfer Accuracy

**Automatic Evaluation:** We measure the accuracy of style transfer for the generated sentences using a pre-trained style classifier. The classifier is trained on data that is not used for

<sup>8</sup>Alternatively, we could use a pivot language that is typologically more distant from English, e.g., Chinese. In this case we hypothesize that stylistic traits would be even less preserved in translation, but the quality of back-translated sentences would be worse. We have not yet investigated how the accuracy of the translation model, nor the language of translation affects our models.

	Politeness				Gender				Political			
	Acc	BL-s	MET	ROU	Acc	BL-s	MET	ROU	Acc	BL-s	MET	ROU
CAE	<b>99.62</b>	6.94	10.73	25.71	65.21	9.25	14.72	42.42	77.71	3.17	7.79	27.17
BST	60.75	2.55	9.19	18.99	54.4	20.73	22.57	55.55	<b>88.49</b>	10.71	16.26	41.02
DRG	90.25	11.83	18.07	41.09	36.29	22.9	22.84	53.30	69.79	25.69	21.6	51.8
TaG	89.50	<b>70.44</b>	<b>36.26</b>	<b>70.99</b>	<b>82.21</b>	<b>52.76</b>	<b>37.42</b>	<b>74.59</b>	87.74	<b>68.44</b>	<b>45.44</b>	<b>77.51</b>

TABLE 3.5: Results on the Politeness, Gender and Political datasets.

	Sentiment				
	Acc	BL-s	BL-r	MET	ROU
CAE	72.1	19.95	7.75	21.70	55.9
DRG	<b>88.8</b>	36.69	14.51	32.09	61.06
TaG	86.6	<b>47.14</b>	<b>19.76</b>	<b>36.26</b>	<b>70.99</b>

TABLE 3.6: Results on the sentiment modification task on Yelp dataset.

training the style transfer generative models shown in Table 3.1. We transfer the style of test sentences and then test the classification accuracy of the generated sentences for the opposite label. For example, if we want to transfer the style of male Yelp reviews to female, then we use the fixed common encoder of the back-translation model to encode the test male sentences and then we use the female generative model to generate the female-styled reviews. We then test these generated sentences for the *female* label using the gender classifier.

The classifier has an accuracy of 82% for the gender-annotated corpus, 92% accuracy for the political slant dataset and 93.23% accuracy for the sentiment dataset.

In Table 3.4, we detail the accuracy of each classifier on generated style-transferred sentences.<sup>9</sup> On two out of three tasks our model substantially outperforms the baseline, by up to 12% in political slant transfer, and by up to 7% in sentiment modification. Table 3.5 shows that the classifier accuracy on the generations of TaG model are comparable (within 1%) with that of DRG for the Politeness dataset. TaG model performs the best in transfer accuracy for the gender transfer task and performs comparable to the BST model on the political slant transfer task. It also performs comparable to the DRG model on the sentiment modification task as shown in Table 3.6.

**Human Evaluation:** Li et al. (2018a) introduce human evaluation for assessing the strength of transfer. Human judges are asked to annotate the generated sentence on a scale of 1 to 5 for similarity to target attribute. Although this is a good practice, demographic attributes such as gender, age and personal choices such political slant etc must not be evaluated by human judgements as there is a danger of bias and stereotypes introduced by people during the evaluation process. This work has performed an analysis of the correlation of the human

<sup>9</sup>In each experiment, we report aggregated results across directions of style transfer; same results broke-down to style categories are listed in the Supplementary Material.



	Content		Attribute		Grammar	
	DRG	TaG	DRG	TaG	DRG	TaG
Politeness	2.9	<b>3.6</b>	3.2	<b>3.6</b>	2.0	<b>3.7</b>
Gender	3.0	<b>3.5</b>	-	-	2.2	<b>2.5</b>
Political	2.9	<b>3.2</b>	-	-	2.5	<b>2.7</b>
Sentiment	3.0	<b>3.7</b>	3	<b>3.9</b>	2.7	<b>3.3</b>

TABLE 3.7: Human evaluation on Politeness, Gender, Political and Yelp datasets.

judgements with the automatic evaluation and argues that it depends on the dataset and the task. Hence, the correlation cannot be taken for granted.

The same instructions from Li et al. (2018a) is used for human evaluation of target attribute match. Overall, both systems (DRG and TaG) are evaluated on a total of 200 samples for Politeness and 100 samples for Yelp. Table 3.7 shows that TaG performs significantly better than DRG on the target attribute matching metric.

### 3.3.2 Preservation of Meaning

**Automatic Evaluation:** To measure preservation of meaning in style transfer, some works have borrowed metrics from other generation or translation tasks such as BLEU (Papineni et al., 2002), ROUGE (Lin and Hovy, 2002) or METEOR (Denkowski and Lavie, 2011). Li et al. (2018a) have released a test set of human references primarily for the sentiment modification task. In this case, you can calculate BLEU between the human references and the generated sentences. In cases where the human references are not available, BLEU is calculated between the generated sentence and the input sentence (referred to as BLEU-s in Table 3.5).

Table 3.5 shows that TaG model achieves significantly higher scores on BLEU, ROUGE and METEOR as compared to the baselines DRG, CAE and BST on the Politeness, Gender, and Political datasets. The BLEU score on the Politeness task is greater by 58.61 points with respect to DRG. In general, CAE and BST achieve high classifier accuracies but they fail to retain the original content.

Table 3.6 compares TaG model against CAE and DRG on the Yelp dataset for sentiment modification task. The test set comprises 500 samples (with human references) curated by Li et al. (2018a). We observe an increase in the BLEU-reference scores by 5.25 on the sentiment modification task.

**Human Evaluation:** Meaning preservation in style transfer is not trivial to define as literal meaning is likely to change when style transfer occurs. For example “My girlfriend loved the desserts” vs “My partner liked the desserts”. Thus we must relax the condition of literal meaning to *intent* or *affect* of the utterance within the context of the discourse. Thus if the

Experiment	CAE	No Pref.	BST
Gender	15.23	41.36	<b>43.41</b>
Political slant	14.55	<b>45.90</b>	39.55
Sentiment	35.91	<b>40.91</b>	23.18

TABLE 3.8: Human preference for meaning preservation in percentages.

intent is to criticize a restaurant’s service in a review, changing “salad” to “chicken” could still have the same effect but if the intent is to order food that substitution would not be acceptable. Ideally we wish to evaluate transfer within some downstream task and ensure that the task has the same outcome even after style transfer. This is a hard evaluation and hence we resort to a simpler evaluation of the “meaning” of the sentence.

We set up a manual pairwise comparison following [Bennett \(2005\)](#). The test presents the original sentence and then, in random order, its corresponding sentences produced by the baseline and our models. For the gender style transfer we asked “Which transferred sentence maintains the same sentiment of the source sentence in the same semantic context (i.e. you can ignore if food items are changed)”. For the task of changing the political slant, we asked “Which transferred sentence maintains the same semantic intent of the source sentence while changing the political position”. For the task of sentiment transfer we have followed the annotation instruction in ([Shen et al., 2017](#)) and asked “Which transferred sentence is semantically equivalent to the source sentence with an opposite sentiment”

We then count the preferences of the eleven participants, measuring the relative acceptance of the generated sentences.<sup>10</sup> A third option “=” was given to participants to mark no preference for either of the generated sentence. The “no preference” option includes choices both are equally bad and both are equally good. We conducted three tests one for each type of experiment - gender, political slant and sentiment. We also divided our annotation set into short ( $\#tokens \leq 15$ ) and long ( $15 < \#tokens \leq 30$ ) sentences for the gender and the political slant experiment. In each set we had 20 random samples for each type of style transfer. In total we had 100 sentences to be annotated. Note that we did not ask about appropriateness of the style transfer in this test, or fluency of outputs, only about meaning preservation.

The results of human evaluation between the CAE and BST models are presented in [Table 3.8](#). Although a no-preference option was chosen often—showing that state-of-the-art systems are still not on par with human expectations—the BST models outperform the baselines in the gender and the political slant transfer tasks.

Crucially, the BST models significantly outperform the CAE models when transferring style in longer and harder sentences. Annotators preferred the CAE model only for 12.5% of the long sentences, compared to 47.27% preference for the BST model.

<sup>10</sup>None of the human judges are authors of this paper

For a fair comparison of the DRG and TaG model, the same instructions from Li et al. (2018a) were used for the human study on content preservation. The reviewers give a score between 1-5 to each of the outputs, where 1 reflects a poor performance on content preservation and 5 means a high content preservation. Table 3.7 shows the results of human evaluations between the DRG and TaG models. We observe a significant improvement in content preservation scores across all datasets (specifically in Politeness domain) highlighting the ability of our model to retain content better than DRG.

### 3.3.3 Fluency

**Automatic Evaluation:** Yang et al. (2018b); He et al. (2020); Lample et al. (2018) use perplexity to measure the fluency of the generated sentences. In most cases perplexity is not correlated with human judgements of fluency.

**Human Evaluation:** We evaluate the fluency of the sentences generated by CAE and BST models. Fluency was rated from 1 (unreadable) to 4 (perfect) as is described in (Shen et al., 2017). We randomly selected 60 sentences each generated by the baseline and the BST model.

The results shown in Table 3.9 are averaged fluency scores for CAE and BST model.

Experiment	CAE	BST
Gender	2.42	<b>2.81</b>
Political slant	2.79	<b>2.87</b>
Sentiment	3.09	<b>3.18</b>
Overall	2.70	<b>2.91</b>
Overall Short	3.05	<b>3.11</b>
Overall Long	2.18	<b>2.62</b>

TABLE 3.9: Fluency of the generated sentences.

BST outperforms the baseline overall. It is interesting to note that BST generates significantly more fluent longer sentences than the baseline model. Since the average length of sentences was higher for the gender experiment, BST notably outperformed the baseline in this task, relatively to the sentiment task where the sentences are shorter.

Table 3.7 shows the results for DRG and TaG model for fluency or grammaticality of the generated content. We observe that TaG model performs much better than DRG on all four tasks for fluency metric.

**Discussion:** For the BST model, the loss function of the generators given in Eq. 3.5 includes two competing terms, one to improve meaning preservation and the other to improve the style transfer accuracy. In the task of sentiment modification, the BST model preserved meaning

worse than the baseline, on the expense of being better at style transfer. We note, however, that the sentiment modification task is not particularly well-suited for evaluating style transfer: it is particularly hard (if not impossible) to disentangle the sentiment of a sentence from its propositional content, and to modify sentiment while preserving meaning or intent. On the other hand, the style-transfer accuracy for gender is lower for BST model but the preservation of meaning is much better for the BST model, compared to CAE model and to “No preference” option. This means that the BST model does better job at closely representing the input sentence while taking a mild hit in the style transfer accuracy.<sup>11</sup>

While popular, the metrics of Transfer Accuracy and BLEU have significant shortcomings making them susceptible to simple adversaries. BLEU relies heavily on n-gram overlap and classifiers can be fooled by certain polarizing keywords. We test this hypothesis on the sentiment transfer task by a *Naive Baseline*. This baseline adds “*but overall it sucked*” at the end of the sentence to transfer it to negative sentiment. Similarly, it appends “*but overall it was perfect*” for transfer into a positive sentiment. This baseline achieves an average accuracy score of 91.3% and a BLEU score of 61.44 on the Yelp dataset. Despite the stellar performance, it does not reflect a high rate of success on the task. In summary, evaluation via automatic metrics might not truly correlate with task success.

### 3.3.4 Manual Inspection

Input	DRG Output	Our Model Output	Strategy
what happened to my personal station?	what happened to my mother to my co???	could you please let me know what happened to my personal station?	Counterfactual Modal
yes, go ahead and remove it.	yes, please go to the link below and delete it.	yes, we can go ahead and remove it.	1st Person Plural
not yet-i’ll try this wk-end.	not yet to say-i think this will be a <unk> long.	sorry not yet-i’ll try to make sure this wk	Apologizing
please check on metro-media energy,	thanks again on the energy industry,	please check on metro-media energy, thanks	Mitigating please start

TABLE 3.10: Qualitative Examples comparing the outputs from DRG and Our model for the Politeness Transfer Task

We compare the results of TaG model with the DRG model qualitatively as shown in Table 3.10. Our analysis is based on the linguistic strategies for politeness as described in (Danescu-Niculescu-Mizil et al., 2013). The first sentence presents a simple example of the *counterfactual modal* strategy inducing “*Could you please*” to make the sentence polite. The second sentence highlights another subtle concept of politeness of *1st Person Plural* where adding “*we*” helps being indirect and creates the sense that the burden of the request is shared between speaker and addressee. The third sentence highlights the ability of the model to add *Apologizing* words like

<sup>11</sup>Details about hyper-paramters, generated examples and additional experiments are provided in Appendix A.

“Sorry” which helps in deflecting the social threat of the request by attuning to the imposition. According to the *Please Start* strategy, it is more direct and insincere to start a sentence with “Please”. The fourth sentence projects the case where our model uses “thanks” at the end to express gratitude and in turn, makes the sentence more polite. TaG model follows the strategies prescribed in (Danescu-Niculescu-Mizil et al., 2013) while generating polite sentences.<sup>12</sup>

**Ablations:** We provide a comparison of the two variants of the tagger, namely the replace-tagger and add-tagger on two datasets for the *Tag and Generate* approach. We also train and compare them with a *combined* variant.<sup>13</sup> We train these tagger variants on the Yelp and Captions datasets and present the results in Table 3.11. We observe that for Captions, where we transfer a factual (neutral) to romantic/humorous sentence, the add-tagger provides the best accuracy with a relatively negligible drop in BLEU scores. On the contrary, for Yelp, where both polarities are clearly defined, the replace-tagger gives the best performance. Interestingly, the accuracy of the add-tagger is  $\approx 50\%$  in the case of Yelp, since adding negative words to a positive sentence or vice-versa neutralizes the classifier scores. Thus, we can use the add-tagger variant for transfer from a polarized class to a neutral class as well.

To check if the combined tagger is learning to perform the operation that is more suitable for a dataset, we calculate the fraction of times the combined tagger performs add/replace operations on the Yelp and Captions datasets. We find that for Yelp (a polar dataset) the combined tagger performs 20% more replace operations (as compared to add operations). In contrast, on the CAPTIONS dataset, it performs 50% more add operations. While the combined tagger learns to use the optimal tagging operation to some extent, a deeper understanding of this phenomenon is an interesting future topic for research. We conclude that the choice of the tagger variant is dependent on the characteristics of the underlying transfer task.

	Sentiment		Captions	
	Acc	BL-r	Acc	BL-r
Add-Tagger	53.2	20.66	<b>93.17</b>	15.63
Replace-Tagger	<b>86.6</b>	19.76	84.5	15.04
Combined	72.5	<b>22.46</b>	82.17	<b>18.51</b>

TABLE 3.11: Comparison of different *tagger* variants for Yelp and Captions datasets

**Changing Content Words:** Given that TaG model is explicitly trained to generate new content only in place of the [TAG] token, it is expected that a well-trained system will retain most of the non-tagged (content) words. Clearly, replacing content words is not desired since it may drastically change the meaning. In order to quantify this, the fraction of non-tagged words being changed across the datasets is computed. The non-tagged words were changed for only

<sup>12</sup>We provide additional qualitative examples for other tasks in the supplementary material.

<sup>13</sup>Training of combined variant is done by training the tagger model on the concatenation of training data for add-tagger and replace-tagger.

6.9% of the sentences. In some of these cases, changing non-tagged words helped in producing outputs that were more natural and fluent.

## 3.4 Related Work

### 3.4.1 Task

Non-parallel style transfer has been largely studied in the field of computer vision (Gatys et al., 2016; Zhu et al., 2017; Luan et al., 2017; Song et al., 2019). The task entails extracting content and style features from a source image, and then synthesizing a new image by combining “content” features of one image with “style” features from another. In natural language processing, the widely studied generation tasks are machine translation and summarization which are trained using parallel sentences. The task of style transfer in text does not typically have parallel sentences (Reddy and Knight, 2016; Voigt et al., 2018; Shen et al., 2017).

Politeness and its close relation with power dynamics and social interactions has been well documented (Brown et al., 1987). Recent work (Danescu-Niculescu-Mizil et al., 2013) in computational linguistics has provided a corpus of *requests* annotated for politeness curated from Wikipedia and StackExchange. Niu and Bansal (2018b) uses this corpus to generate polite dialogues. Their work focuses on contextual dialogue response generation as opposed to content preserving style transfer, while the latter is the central theme of our work. Prior work on Enron corpus (Yeh and Harnly, 2006) has been mostly from a socio-linguistic perspective to observe social power dynamics (Bramsen et al., 2011; McCallum et al., 2007), formality (Peterson et al., 2011) and politeness (Prabhakaran et al., 2014). We build upon this body of work by using this corpus as a source for the style transfer task.

### 3.4.2 Methodology

Style transfer with non-parallel text corpus has become an active research area due to the recent advances in text generation tasks. Hu et al. (2017) use variational auto-encoders with a discriminator to generate sentences with controllable attributes. The method learns a disentangled latent representation and generates a sentence from it using a code. This paper mainly focuses on sentiment and tense for style transfer attributes. It evaluates the transfer strength of the generated sentences but does not evaluate the extent of preservation of meaning in the generated sentences. In our work, we show a qualitative evaluation of meaning preservation.

Shen et al. (2017) first present a theoretical analysis of style transfer in text using non-parallel corpus. The paper then proposes a novel cross-alignment auto-encoders with discriminators architecture to generate sentences. It mainly focuses on sentiment and word decipherment for style transfer experiments.

Fu et al. (2018) explore two models for style transfer. The first approach uses multiple decoders for each type of style. In the second approach, style embeddings are used to augment the encoded representations, so that only one decoder needs to be learned to generate outputs in different styles. Style transfer is evaluated on scientific paper titles and newspaper titles, and sentiment in reviews. This method is different from ours in that we use machine translation to create a strong latent state from which multiple decoders can be trained for each style. We also propose a different human evaluation scheme.

Compared to prior work, “Delete, Retrieve and Generate” (Li et al., 2018a) (referred to as DRG henceforth) and its extension (Sudhakar et al., 2019) are effective methods to generate outputs in the target style while having a relatively high rate of source content preservation. However, DRG has several limitations: (1) the delete module often marks content words as stylistic markers and deletes them, (2) the retrieve step relies on the presence of similar content in both the source and target styles, (3) the retrieve step is time consuming for large datasets, (4) the pipeline makes the assumption that style can be transferred by deleting stylistic markers and replacing them with target style phrases, (5) the method relies on a fixed corpus of style attribute markers, and is thus limited in its ability to generalize to unseen data during test time. *Tag and Generate* methodology differs from these works as it does not require the retrieve stage and makes no assumptions on the existence of similar content phrases in both the styles. This also makes the pipeline faster in addition to being robust to noise.

Wu et al. (2019) treats style transfer as a conditional language modelling task. It focuses only on sentiment modification, treating it as a cloze form task of filling in the appropriate words in the target sentiment. In contrast, TaG is capable of generating the entire sentence in the target style. Further, it is more generalizable and we show results on three other style transfer tasks.

Our work is also closely-related to a problem of paraphrase generation (Madnani and Dorr, 2010; Dong et al., 2017), including methods relying on (phrase-based) back-translation (Ganitkevitch et al., 2011; Ganitkevitch and Callison-Burch, 2014). More recently, Mallinson et al. (2017) and Wieting et al. (2017) showed how neural back-translation can be used to generate paraphrases. An additional related line of research is machine translation with non-parallel data. Lample et al. (2018) and Artetxe et al. (2018) have proposed sophisticated methods for unsupervised machine translation. These methods could in principle be used for style transfer as well.

### 3.5 Conclusion

This chapter begins by outlining the *style* aspect of human communication and what it means. It then defines the task of style transfer and sketches the challenges involved in the task. The political slant task is a novel task that we introduce. We introduce the task of politeness transfer for which we provide a dataset comprised of sentences curated from email exchanges present in the Enron corpus.

We propose two novel approaches to the task of style transfer with non-parallel text. We learn a latent content representation using machine translation techniques; this aids grounding the meaning of the sentences, as well as weakening the style attributes. We apply this technique to three different style transfer tasks. In transfer of political slant and sentiment we outperform an off-the-shelf state-of-the-art baseline using a cross-aligned autoencoder. Our model also outperforms the baseline in all the experiments of fluency, and in the experiments for meaning preservation in generated sentences of gender and political slant. Yet, we acknowledge that the generated sentences do not always adequately preserve meaning.

We extend prior works on style transfer by introducing a simple pipeline - tag and generate which is an interpretable two-staged approach for content preserving style transfer. We believe our approach is the first to be robust in cases when the source is style neutral, like the "non-polite" class in the case of politeness transfer. Automatic and human evaluation shows that our approach outperforms other state-of-the-art models on content preservation metrics while retaining (or in some cases improving) the transfer accuracies.

These techniques are suitable not just for style transfer, but for enforcing style, and removing style too. Future work can apply this technique to debiasing sentences and anonymization of author traits such as gender and age.

Measuring the separation of style from content is hard, even for humans. It depends on the task and the context of the utterance within its discourse. Ultimately we must evaluate our style transfer within some down-stream task where our style transfer has its intended use but we achieve the same task completion criteria.



## Chapter 4

# Document Grounded Generation

### Monkey selfie copyright dispute

From Wikipedia, the free encyclopedia

The monkey selfie copyright dispute is a series of disputes about the copyright status of selfies by macaques.

On 4 July 2011 several publications picked up the story and quoted Slater as describing the photographs as self-portraits. Slater said reports that a monkey ran off with his camera and "began taking self-portraits" were incorrect and that the portrait was shot when his camera had been on a tripod, with the primates playing around with a remote cable release as he fended off other monkeys.<sup>[14]</sup>



One of the monkey selfies at issue in the dispute

#### Ape-rture priority photographer plays down monkey reports

amateurphotographer.co.uk  
July 5, 2011

**A photographer who says he witnessed monkeys taking pictures of themselves, tells Amateur Photographer (AP) that much of the media coverage has been exaggerated.**

Speaking to AP, David explained that his camera had been mounted on a tripod when the primates began playing around with a remote 'cable release' as he was trying to fend off other monkeys.

FIGURE 4.1: Example of content transfer: Given existing context (yellow) and a document with additional relevant information (green), the task is to update the context (orange) to reflect the most salient updates.

Natural language generation (NLG) systems are increasingly expected to be naturalistic, content-rich, and situation-aware due to their popularity and pervasiveness in human life (Reiter and Dale, 2000; Mitchell et al., 2014). Recent work in neural natural language generation (NLG) has witnessed a growing interest in controlling text for various form-related and linguistic properties, such as style (Ficler and Goldberg, 2017), affect (Ghosh et al., 2017), politeness (Sennrich et al., 2016), persona (Li et al., 2016b) voice (Yamagishi et al., 2016), grammatical correctness

(Ji et al., 2017), and length (Kikuchi et al., 2016). This trend offers the promise of empowering existing authoring tools such as Grammarly, Google Smart Compose, and Microsoft Word with the ability to control a much greater variety of textual properties, which are currently mostly limited to grammar, spelling, word choice, and wordiness. What has been relatively less explored in neural NLG research is the ability to control the generation of a current sentence not only in its *form*, but also its *content*. This is particularly relevant in dialogue systems (Zhang et al., 2018; Niu and Bansal, 2018b), machine translation systems (Mirkin and Meunier, 2015; Rabinovich et al., 2017), story generation (Fan et al., 2018; Yao et al., 2019), and question answering systems (Gatus, 2017; Reddy et al., 2019).

Despite these mainstream applications, NLG systems face the challenges of being bland, devoid of content, generating generic outputs and hallucinating information (Wiseman et al., 2017; Li et al., 2016a; Holtzman et al., 2020; Welleck et al., 2020). Grounding the generation in different modalities like images (Huang et al., 2016; Mostafazadeh et al., 2017; Shuster et al., 2018), videos (Palaskar et al., 2019; Regneri et al., 2013), and structured data (Banik et al., 2013; Gardent et al., 2017) alleviates some of these issues. Generating natural language from schematized or structured data such as database records, slot-value pair, and Wikipedia Infobox has been explored in prior work (Mei et al., 2016; Wen et al., 2015; Lebrecht et al., 2016). Although useful, these tasks encounter difficulties such as general applicability (databases may not be available for all domains) and are constrained by the available resources (size of the database).

Document grounded generation mitigates these applicability issues by exploiting the vast availability of data in unstructured form (e.g. books, encyclopedias, news articles, and Wikipedia articles). This enhances the applicability of document grounded generation to a wide range of domains with limited (or no) availability of structured data. Hence, recent work has focused on defining new tasks and carving the scope of the problems (Liu et al., 2018; Prabhumoye et al., 2019b; Faltings et al., 2020; Zhou et al., 2018; Dinan et al., 2018).

Consider for example Figure 4.1, which illustrates a situation where an author edits a document (here a Wikipedia article), and the goal is to generate or suggest a next sentence (shown in orange) to the author. This type of unconstrained, long-form text generation task (Mostafazadeh et al., 2016; Fan et al., 2018) is of course extremely difficult. Free-form generation can easily go astray due to two opposing factors. On one hand, ensuring that the generated output is of relatively good quality often comes at the cost of making it bland and devoid of factual content (Li et al., 2016a). On the other hand, existing techniques can help steer neural models away from blandness in order to produce more contentful outputs (using temperature sampling (Fan et al., 2018), GAN (Goodfellow et al., 2014), etc.), but often at the cost of “hallucinating” (Wiseman et al., 2017) words or concepts that are totally irrelevant. Neither situation provides a compelling experience to the user.

What is clearly missing from the aforementioned authoring scenario is the notion of *grounding*: there is often a profusion of online resources that bear at least some relevance to any given document currently being written. Much of the general-purpose world knowledge is available in the form of encyclopedias (e.g., Wikipedia), books (e.g., Project Gutenberg, Google Books),

**User1:** The Notebook is hands-down one of my favorite movies EVER! Have you ever seen The Notebook?

**User2:** No I have never seen this movie. I am going to try it out now

**User1:** It was a heartwarming story of young love. The main characters are played by Ryan Gosling and Rachel McAdams.

**User2:** Ok this sounds nice. I think Ryan is a good actor.

**User1:** For all the praise it received, I was surprised to see that it only got a 5.7/10 on Rotten Tomatoes.

**User2:** That is interesting. They never get the rating correct.

**User1:** Ryan is a great actor, as well as Rachel McAdams. The story goes back and forth between present day and the past. Older Ryan is played by James Garner and older Rachel is played by Gena Rowlands. Yeah, Rotten Tomatoes never gets the right ratings..LOL. I always like to see the ratings but if I want to see a movie, I will watch it even if it has a bad rating.

### The Notebook

From Wikipedia, the free encyclopedia

For other uses, see [Notebook \(disambiguation\)](#).

**The Notebook** is a 2004 romantic drama film directed by Nick Cassavetes and written by Jeremy Leven from Jan Sardi's adaptation of the 1996 novel by Nicholas Sparks. The film stars Ryan Gosling and Rachel McAdams as a young couple who fall in love in the 1940s. Their story is narrated from the present day by an elderly man (played by James Garner), telling the tale to a fellow nursing home resident (played by Gena Rowlands, who is Cassavetes's mother).

The Notebook received generally mixed reviews, but performed well at the box office and received a number of award nominations, winning eight Teen Choice Awards, a Satellite Award, and an MTV Movie Award. The film became a sleeper hit<sup>[R]</sup> and has gained a cult following.<sup>[R]</sup> On November 11, 2012, ABC Family premiered an extended version with deleted scenes added back into the original storyline.<sup>[7]</sup>

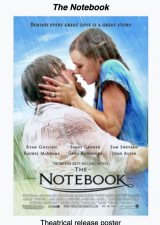
**Reception**  [[edit](#)]

**Box office**  [[edit](#)]

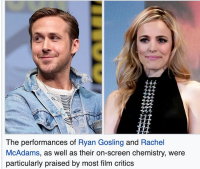
The film premiered June 25, 2004, in the United States and Canada and grossed \$13,464,745 in 2,303 theaters its opening weekend, ranking number 4 at the box office.<sup>[8]</sup> The film grossed a total of \$115,603,229 worldwide, \$81,001,787 in Canada and the United States and \$34,601,442 in other countries.<sup>[2]</sup> It is the 15th highest-grossing romantic drama film of all time.<sup>[24]</sup>

**Critical reception**  [[edit](#)]

The Notebook received a mixed reaction from film critics. The 178 reviews on review aggregator Rotten Tomatoes show that 53% of critics gave the film a positive review, with an average rating of 5.64/10 and the website's consensus stating "It's hard not to admire its unabashed sentimentality, but *The Notebook* is too clumsily manipulative to rise above its melodramatic clichés."<sup>[9]</sup> At *Metacritic*, which assigns an average rating out of 100 to reviews from mainstream critics, the film currently holds an average score of 53, based on 34 reviews, which indicates "mixed or average reviews."<sup>[25]</sup>



Theatrical release poster  
Directed by Nick Cassavetes



The performances of Ryan Gosling and Rachel McAdams, as well as their on-screen chemistry, were particularly praised by most film critics

FIGURE 4.2: Example of human-human dialogue where *User1* has access to the Wikipedia document and *User2* does not. The information underlined in red is taken from the Wikipedia article by *User1*.

and news articles. While the generation of good quality texts without any conditioning on “external” sources (Fan et al., 2018) might be an interesting research endeavor on its own, we argue that grounding can make the generation task much easier, e.g., as shown in Figure 4.1 where a passage of a news article (green) can be reformulated considering the current context of the document (yellow) in order to produce a natural next sentence (orange). In light of this desideratum, this chapter addresses the problem of grounded text generation, where the goal is to infuse the content or knowledge from an external unstructured source (e.g., a news article as in Figure 4.1) in order to generate a follow-up sentence of an existing document. We see this as a form of *Content Transfer*, as other characteristics of the external source—such as style and linguistic form—are not controlled.

Apart from the aforementioned scenario, generation grounded in an external unstructured source of information is very useful in various other scenarios like generating factual dialogue responses from a given document, generating stories based on a plot, generating one coherent report from multiple source documents as in the case of scientific summary etc. Particularly, we are also interested in dialogue response generation. Most of the dialog systems hallucinate a response given the context. We introduce a new dataset with human-human conversations which grounds the dialogue responses in Wikipedia articles about a topic. In this case, the current context is the current dialogue history, and We are interested in generating the appropriate response from the information in the external document (Wikipedia article). Figure 4.2 shows an example of this task from the dataset collected. Although the dataset is based on the topic on movies, we see the same techniques being valid for other external documents such as manuals, instruction booklets, and other informational documents.

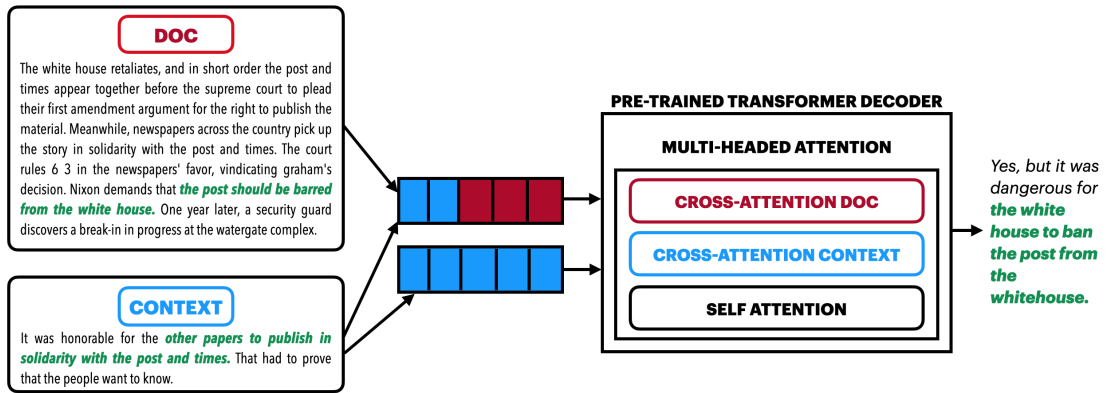


FIGURE 4.3: Document Grounded Generation: An example of a conversation that is grounded in the given document (text in green shows information from the document that was used to generate the response).

**Overview:** We first formally define the general task of document grounded generation in §4.1.1. We then discuss the task of Wikipedia Update Generation and the process of data curation in §4.1.2. This work was done in collaboration with Chris Quirk and Michel Galley from Microsoft Research. The task of document grounded dialogue response generation along with the data collection process is described in §4.1.3. This work was done in collaboration with Kangyan Zhou and Alan W Black. We describe the RNN-based generative models in §4.2.1, extractive models in §4.2.2 and extensions of pre-trained encoder decoder models in §4.2.3. The experiments and results are presented in §4.3. A comprehensive manual inspection is showcased in §4.3.3. A literature survey is presented in §4.5.

## 4.1 Tasks and Datasets

### 4.1.1 Task Definition

Our task is to generate text given a context and a source of content (document). Additionally, the generated text should coherently fit the context and contain information from the document. We focus on content present in unstructured form in documents to ground text generation. Formally we define our task as follows: given an existing context (or curated text)  $s$  and a document  $d$  describing novel information relevant to the context, the system must produce a revised text  $s'$  that incorporates the most salient information from  $d$ . We restrict our focus to the cases where the revised text  $s'$  can be obtained by appending the new information from  $d$  to the original context  $s$ .<sup>1</sup> In particular, we assume that we can transform the context  $s$  into the new text  $s'$  by appending one additional update sentence  $x$  to  $s$ . This makes the same techniques applicable to the dialogue response generation task.

<sup>1</sup>In the case of generating Wikipedia Updates and similar tasks, updated information from  $d$  might demand substantial changes to  $s$ : perhaps core assumptions of  $s$  were contradicted, necessitating many removed and rewritten sentences. We postpone this complex setting to future work.

Figure 4.3 illustrates an example of the dialogue response generation task. Dialogue response generation is traditionally conditioned on the dialogue context (Vinyals and Le, 2015; Li et al., 2016a). As Figure 4.3 demonstrates, the generative model is conditioned on both the document as well as the dialogue context. Note that the context and document play different roles in impacting the generation – the context sets the background while the document provides the content necessary to generate the text.

Formally, each sample  $\mathbf{i}$  of our task is defined as a tuple  $(\mathbf{d}_i, \mathbf{s}_i, \mathbf{x}_i)$  containing context  $\mathbf{s}_i$ , document  $\mathbf{d}_i$  and text  $\mathbf{x}_i$  to be generated. Note that each  $\mathbf{d}_i$  can be a single document or a set of documents. The task is to generate  $\mathbf{x}_i$  such that it coherently follows  $\mathbf{s}_i$  and contains information from  $\mathbf{d}_i$ . The task can be modeled as the following conditional text generation model:

$$p_{\theta}(\mathbf{x}_i | \mathbf{s}_i, \mathbf{d}_i),$$

where  $\theta$  is a set of model parameters.

Figure 4.3 illustrates that the generator has to account for two inputs the dialogue context  $\mathbf{s}_i$  (shown in blue) and the document  $\mathbf{d}_i$  (shown in red) to generate the response  $\mathbf{x}_i$  grounded in  $\mathbf{d}_i$  (text shown in green). If the generative model was only conditioned on dialogue context, then it could produce generic responses like “Do you think they did the right thing?” or “Yes, I agree.” or hallucinate information like “Yes, and the Times published it on the front page.”. These which would be appropriate to the given context but are devoid of content or contain wrong information. Document grounded models are capable of responding with interesting facts like “Yes, but it was dangerous for the white house to ban the post from the white house.”

We discuss the task of Wikipedia update generation as well as the dataset collected to explore this task in §4.1.2. In §4.1.3, we discuss the task of document grounded dialogue response generation and describe in detail the collection of CMU Document Grounded Conversations Dataset (CMU\_DoG) (Zhou et al., 2018). We also briefly describe the Wizard of Wikipedia dataset (Dinan et al., 2018) in §4.1.3.

### 4.1.2 Wikipedia Update Generation

This task involves generating an update for Wikipedia context given a news article (Prabh-moye et al., 2019b). It consists tuples of the form  $(\mathbf{d}_i, \mathbf{s}_i, \mathbf{x}_i)$ , where the grounding document  $\mathbf{d}_i$  is the news article which contains information for the reference update  $\mathbf{x}_i$ .  $\mathbf{x}_i$  is written by a Wikipedia editor as an update to the Wikipedia context  $\mathbf{s}_i$ . The goal of the task is to generate  $\mathbf{x}_i$  given the context  $\mathbf{s}_i$  and the document  $\mathbf{d}_i$ .

Wikipedia can provide a naturally-occurring body of text with references to primary sources. A substantial fraction of Wikipedia sentences include citations to supporting documentation, a ripe source of data for content transfer. That said, some of the citations are quite difficult to follow or trust: broken URLs might lead to lost information; citations to books are difficult to

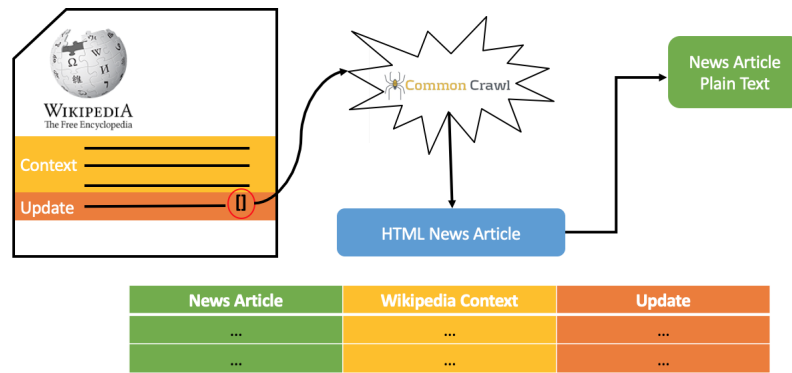


FIGURE 4.4: Dataset creation process for Wikipedia Edit Generation

consume given the large scope of information; etc. Therefore, cases where the reference links to some well-known news sources are considered.

Based on citation frequency, we selected a list of 86 domains,<sup>2</sup> primarily news outlets. During the data creation process we only considered citations belonging to one of these eighty six domains. This simplifying assumption is made for several reasons. First, our English Wikipedia dump contained approximately 23.7 million citation URLs belonging to 1.6 million domains; fine-grained filtering would be a daunting task. The hand-vetted list of domains is a high-precision (albeit low-recall) means of selecting clean data. Second, we wanted to ground the generated text on credible, consistent, and well-written sources of information. Furthermore, well-known domains are readily available on Common Crawl,<sup>3</sup> leading to an easily-reproducible dataset.

Figure 4.4 illustrates the procedure used to create a dataset for the Wikipedia Edit generation task shown in Figure 4.1. For each Wikipedia article, we extracted the plain text without markdown. When encountering a citation belonging to a selected domain, we considered the sentence just before the citation to be generated based on the content of the citation. This sentence became our reference update sentence: the additional update sentence  $x$  added to the context  $s$  to produce the new text  $s'$ . The  $k$  sentences prior to the target sentence in the Wikipedia article were considered to be the curated text  $s$ . In this case, we used a window of  $k = 3$  sentences to select the context. The cited article acted as the document  $d$ , from which the appropriate update  $x$  can be generated.

The HTML source of the citation was downloaded from Common Crawl for reproducibility and consistency. The HTML derived from Common Crawl is then processed to get the plain text of the news article. The resulting dataset  $\mathcal{C}$  consists of aligned tuples  $\mathcal{C} = (d_i, s_i, x_i)_{i \in [1, n]}$ , where  $n$  is the total number of samples in the dataset.

Alternatively, one might rely on Wikipedia edit history to create a dataset. In this setting, edits which include a new citation would act as the update  $x$ . Although this has the upside of

<sup>2</sup>This list is provided in the data release of this paper.

<sup>3</sup><http://commoncrawl.org/>

Corpus	Input	Output	#Examples	Rouge-1 R
Gigaword (Graff and Cieri, 2003)	$10^1$	$10^1$	$10^6$	78.7
CNN/DailyMail (Nallapati et al., 2016)	$10^2-10^3$	$10^1$	$10^5$	76.1
WikiSum (Liu et al., 2018)	$10^2-10^6$	$10^1-10^3$	$10^6$	59.2
Content Transfer (this paper)	$10^1-10^3$	$10^1-10^2$	$10^5$	66.9

TABLE 4.1: Key characteristics of the dataset: approximate size of input and output instances, approximate dataset size, and recall of reference output against the source material, as a measure of dataset difficulty.

identifying potentially complex, multi-sentence updates, preliminary analysis suggested that these edits are noisy. Editors may first generate the content in one edit, then add the citation in a subsequent edit, they may only rephrase a part of the text while adding the citation, or they may check in a range of changes across the document in a single edit. Our simpler sentence-based approach leads to an interesting dataset with fewer complications.

**Dataset Statistics and Analysis:** Table 4.1 describes some key statistics of this dataset and how it compares with other datasets used for similar tasks. The ROUGE-1 recall scores of reference output  $x$  against document  $d$  suggest this task will be difficult for conventional extractive summarization techniques.<sup>4</sup> We hypothesize that during content transfer, the language in document  $d$  often undergoes substantial transformations to fit the context  $s$ . The average unigram overlap (after stopword removal) between the document  $d$  and the reference update sentence  $x$  is 55.79%; overlap of the context  $s$  and the reference update sentence  $x$  is 30.12%. This suggests the reference update sentence  $x$  can be derived from the document  $d$ , though not extracted directly. Furthermore, the content of  $x$  is very different from the content of  $s$  but appears topically related.

Our dataset consists of approximately 290k unique Wikipedia articles. Some heavily-cited articles include ‘Timeline of investigations into Trump and Russia (2017)’, ‘List of England Test cricketers’, and ‘2013 in science’. We randomly split the dataset into 580k training instances, 6049 validation instances, and 50k test instances, ensuring that any Wikipedia article appearing in the train set must not appear in validation or test.

### 4.1.3 Document Grounded Dialog Generation

Goal oriented dialogues have been traditionally grounded in structured sources like slot-value pairs and databases (Wei et al., 2018; Rastogi et al., 2020). Open domain dialogue generation on the other hand faces the issue of “hallucinating” information (Ghazvininejad et al., 2018). Hence, we study open domain dialogue generation which is grounded in documents as a source of information.

<sup>4</sup>ROUGE-1 recall was computed on a sample of 50k instances from the entire dataset.

## CMU Document Grounded Conversations (CMU\_DoG)

The CMU Document Grounded Conversations dataset consists of human-human conversations collected over Amazon Mechanical Turk (Zhou et al., 2018). The conversations are grounded in a document provided to the crowd-workers and focuses only on movies. The dataset uses Wikipedia descriptions of movies for grounding the conversations. The dataset consists tuples of the form  $(\mathbf{d}_i, \mathbf{s}_i, \mathbf{x}_i)$ , where  $\mathbf{d}_i$  is a section (or passage) extracted from Wikipedia,  $\mathbf{s}_i$  is dialogue history (or context) and  $\mathbf{x}_i$  is the reference response. The response  $\mathbf{x}_i$  is grounded in  $\mathbf{d}_i$  and coherently follows the conversation  $\mathbf{s}_i$ . An example conversation from this dataset is shown in Figure 4.2.

To create a dataset for this task, the following were required: (1) A set of documents (2) Two humans chatting about the content of the document for more than 12 turns. We collected conversations about the documents through Amazon Mechanical Turk (AMT) and restricted the topic of the documents to be movie-related articles to facilitate the conversations. Initially, we experimented with different potential domains. Since movies are engaging and widely known, people actually stay on task when discussing them. In fact in order to make the task interesting, we offered a choice of movies to the participants so that they are invested in the task.

### Document Set Creation

We chose Wikipedia<sup>5</sup> articles to create a set of documents  $D = \{d_1, \dots, d_{30}\}$  for grounding of conversations. We randomly selected 30 movies, covering various genres like thriller, superhero, animation, romantic, biopic etc. We extracted the key information provided in the Wiki article and divide it into four separate sections. This was done to reduce the load of the users to read, absorb and discuss the information in the document. Hence, each movie document  $d_i$  consists of four sections  $\{e_1, e_2, e_3, e_4\}$  corresponding to basic information and three key scenes of the movie. The basic information section  $e_1$  contains data from the Wikipedia article in a standard form such as year, genre, director. It also includes a short introduction about the movie, ratings from major review websites, and some critical responses. Each of the key scene sections  $\{e_2, e_3, e_4\}$  contains one short paragraph from the plot of the movie. Each paragraph contains on an average 7 sentences and 143 words. These paragraphs were extracted automatically from the original articles, and were then lightly edited by hand to make them of consistent size and detail.

### Dataset Creation

To create a dataset of conversations which uses the information from the document, involves the participation of two workers. Hence, we explore two scenarios: (1) Only one worker has

---

<sup>5</sup><http://en.wikipedia.org>



---

User 2: Hey have you seen the inception?  
 User 1: No, I have not but have heard of it. What is it about  
 User 2: It's about extractors that perform experiments using military technology on people to retrieve info about their targets.

---

TABLE 4.2: An example conversation for *scenario 1*. User 1 does not have access to the document, while User 2 does.

---

User 1: Hi  
 User 2: Hi  
 User 2: I thought The Shape of Water was one of Del Toro's best works. What about you?  
 User 1: Did you like the movie?  
 User 1: Yes, his style really extended the story.  
 User 2: I agree. He has a way with fantasy elements that really helped this story be truly beautiful.  
 User 2: It has a very high rating on rotten tomatoes, too. I don't always expect that with movies in this genre.

---

TABLE 4.3: An example conversation for *scenario 2*. Both User 1 and User 2 have access to the Wiki document.

access to the document and the other worker does not and (2) Both the workers have access to the document. In both settings, they are given the common instructions of chatting for at least 12 turns.

**Scenario 1: One worker has document.** In this scenario, only one worker has access to the document. The other worker cannot see the document. The instruction to the worker with the document is: *Tell the other user what the movie is, and try to persuade the other user to watch/not to watch the movie using the information in the document*; and the instruction to the worker without the document is: *After you are told the name of the movie, pretend you are interested in watching the movie, and try to gather all the information you need to make a decision whether to watch the movie in the end*. An example of a dialogue for this scenario is shown in Table 4.2.

**Scenario 2: Both workers have document.** In this scenario, both the workers have access to the same Wiki document. The instruction given to the workers are: *Discuss the content in the document with the other user, and show whether you like/dislike the movie*. An example of the dialogue for this scenario is shown in Table 4.3.

**Workflow:** When two workers enter the chat-room, they are given only the first section on basic information  $e_1$  of the document  $d_i$ . After they complete 3 turns (for the first section 6 turns is needed due to initial greetings), the users will be shown the next section. The workers are encouraged to discuss information in the new section, but are not constrained to do so.

Dataset	# Utterances	Avg. # of Turns
CMU-DoG	130,000	31.00
Persona-chat (Zhang et al., 2018)	164,356	14.00
Cornell Movie (Danescu-Niculescu-Mizil and Lee, 2011)	304,713	1.38
Frames dataset (El Asri et al., 2017)	19,986	15.00

TABLE 4.4: Comparison with other datasets. The average number of turns are calculated as the number of utterances divided by the number of conversations for each of the datasets.

**Dataset Statistics:** The dataset consists of total 4112 conversations with an average of 21.43 turns. The number of conversations for *scenario 1* is 2128 and for *scenario 2* it is 1984. We consider a turn to be an exchange between two workers (say  $w_1$  and  $w_2$ ). Hence an exchange of  $w_1, w_2, w_1$  has 2 turns ( $w_1, w_2$ ) and ( $w_2, w_1$ ). We show the comparison of our dataset as **CMU Document Grounded Conversations (CMU-DoG)** with other datasets in Table 4.4. One of the salient features of CMU-DoG dataset is that it has mapping of the conversation turns to each section of the document, which can then be used to model conversation responses. Another useful aspect is that we report the quality of the conversations in terms of how much the conversation adheres to the information in the document.

Percentile	20	40	60	80	99
BLEU	0.09	0.20	0.34	0.53	0.82

TABLE 4.5: The distribution of BLEU score for conversations with more than 10 turns.

**Split Criteria:** We measure the quality of the conversations using BLEU (Papineni et al., 2002) score because we wanted to measure the overlap of the turns of the conversation with the sections of the document. Hence, a good quality conversation should use more information from the document than a low quality conversation. We then divide the dataset into three ratings based on this measure. The BLEU score is calculated between all the utterances  $\{x_1, \dots, x_n\}$  of a conversation  $C_i$  and the document  $d_i$  corresponding to  $C_i$ . Incomplete conversations that have less than 10 turns are eliminated. The percentiles for the remaining conversations are shown in Table 4.5. We split the dataset into three ratings based on BLEU score.

**Rating 1:** Conversations are given a rating of 1 if their BLEU score is less than or equal to 0.1. We consider these conversations to be of low-quality.

**Rating 2:** All the conversations that do not fit in rating 1 and 3 are marked with a rating of 2.

**Rating 3:** Conversations are labeled with a rating of 3, only if the conversation has more than 12 turns and has a BLEU score larger than 0.587. This threshold was calculated by summing

	Rating 1	Rating 2	Rating 3	Rating 2& 3
Total # of conversations	1443	2142	527	2669
Total # of utts	28536	80104	21360	101464
Avg. # utts/conversation	19.77(13.68)	35.39(8.48)	40.53(12.92)	38.01(9.607)
Avg. length of utterance	7.51(50.19)	10.56(8.51)	16.57(15.23)	11.83(10.58)

TABLE 4.6: The statistics of the dataset. Standard deviation in parenthesis.

the mean (0.385) and the standard deviation (0.202) of BLEU scores of the conversations that do not belong rating 1.

The average BLEU score for workers who have access to the document is 0.22 whereas the average BLEU score for the workers without access to the document is 0.03. This suggests that even if the workers had external knowledge about the movie, they have not extensively used it in the conversation. It also suggests that the workers with the document have not used the information from the document verbatim in the conversation. Table 4.6 shows the statistics on the total number of conversations, utterances, and average number of utterances per conversation and average length of utterances for all the three ratings.

**Dataset analysis:** We perform two kinds of automated evaluation to investigate the usefulness of the document in the conversation. The first one is to investigate if the workers use the information from the document  $d_i$  in the conversation. The second analysis is to show that the document adds value to the conversation. Let the set of tokens in the current utterance  $x_i$  be  $\mathbf{N}$ , the set of tokens in the current section  $e_i$  be  $\mathbf{M}$ , the set of tokens in the previous three utterances be  $\mathbf{H}$ , and the set of stop words be  $\mathbf{S}$ . In *scenario 1*, we calculate the set operation *new tokens* as We find that on average 0.78 new tokens (excluding stop words) are introduced in the current utterance  $x_i$  that are present in the current section  $e_i$  but are not present in the prior three utterances. The average length of  $x_i$  is 12.85 tokens. Let the tokens that appear in all the utterances  $(x_i, \dots, x_{i+k})$  corresponding to the current section  $e_i$  be  $\mathbf{K}$  and the tokens that appear in all the utterances  $(x_i, \dots, x_{i+p})$  corresponding to the previous section  $e_{i-1}$  be  $\mathbf{P}$ . In *scenario 2*, we calculate the set operation *new tokens* as we find that on average there are 5.84 common tokens in the utterances that are mapped to the current section  $e_i$  and in  $e_i$  but are not present in the utterances of the previous section  $e_{i-1}$ . The average length of the utterances in a section  $e_i$  is 117.12 tokens. These results show that people use the information in the new sections and are not fixated on old sections. It also shows that they use the information to construct the responses.

## Wizard of Wikipedia

This dataset also consists of human-human conversations collected over Amazon Mechanical Turk and are grounded in passages extracted from Wikipedia (Dinan et al., 2018). These conversations are grounded in a diverse range of topics (totally 1365) which are further split into

seen and unseen topics during training and validation. At each step of the dialogue the wizard has access to a set of passages of knowledge which may be relevant to the given dialogue context. The dataset is created by retrieving the top 7 articles (first paragraph only) that are most relevant to the last two turns of dialogue (by wizard and apprentice). Hence, the dataset consists tuples of the form  $(\mathbf{d}_i, \mathbf{c}_i, \mathbf{x}_i)$ , where  $\mathbf{d}_i$  is a list of 7 passages relevant to the conversation,  $\mathbf{c}_i$  is dialogue history (or context) and  $\mathbf{x}_i$  is the reference response.

## 4.2 Methodology

For training data, we rely on a large dataset of existing context  $\mathbf{S} = \{s_1, \dots, s_n\}$ , corresponding documents with novel information  $\mathbf{D} = \{d_1, \dots, d_n\}$ , and the update sentences  $\mathbf{X} = \{x_1, \dots, x_n\}$ . We have designed the task to generate the update sentence  $x_i$  that could be appended to the context  $s_i$  in order to incorporate the additional information from document  $d_i$ . The goal would be to identify new information (in particular,  $d_i \setminus s_i$ ) that is most salient to the topic or focus of the text, then generate a single sentence that represents this information.

### 4.2.1 Generative models

A natural though difficult means of generating this additional update sentence  $x$  is to use a generative model conditioned on the information in the context  $s$  and the new document  $d$ . Recent methods inspired by successful neural machine translation systems have produced impressive results in abstractive summarization (Nallapati et al., 2016). Hence, our first step is to use the sequence-to-sequence encoder-decoder model (Bahdanau et al., 2015) with attention (Luong et al., 2015b) for our task. This kind of model assumes that the output sentence can be generated word-by-word. Each output word  $x_i^t$  generated is conditioned on all prior words  $x_i^{<t}$  and an encoded representation of the context  $z$ :

$$\prod_t p(\hat{x}_i^t | \hat{x}_i^{<t}, z) \quad (4.1)$$

**Context Agnostic Generative (CAG) Model:** One simple baseline is to train a sequence-to-sequence model for the document  $d$  alone that does not directly incorporate information from the context  $s$ . Here, the algorithm is trained to generate the most likely update sentence  $\hat{x} = \arg \max p(x|d)$ . In this setting, we consider the reference document  $d_i$  as the source and the update sentence to be generated  $x_i$  as the target.

$$z = \text{Encoder}(d_i, \theta) \quad (4.2)$$

The encoder and decoder do not directly see the information from the context  $s$ , but the update  $x$  inherently carries some information about it. The parameters of the model are learned from updates that were authored given the knowledge of the context. Hence, the model may capture some generalizations about the kinds of information and locations in  $d$  that are most likely to contribute novel information to  $s$ .

**Context Only Generative (COG) Model:** This algorithm is trained to generate the most likely update sentence  $\hat{x} = \arg \max p(x|s)$ . This model is similar to CAG except that we consider the context  $s_i$  as the source. In this setting, there is no grounding of the content to be generated.

**Context Informed Generative (CIG) Model:** An obvious next step is to incorporate information from the context  $s$  as well. We can concatenate the document and the context, and produce an encoded representation of this sequence.

$$z = \text{Encoder}([d_i; s_i], \theta) \quad (4.3)$$

This approach incorporates information from both sources, though it does not differentiate them clearly. Thus, the model may struggle to identify which pieces of information are novel with respect to the context. To clearly identify the information that is already present in the context  $s$ , a model could encode  $s$  and  $d$  separately, then incorporate both signals into the generative procedure. This approach makes modification to the external input module (§2.2) of the schema in §2. Specifically, it uses the Arithmetic or Linear Transform technique to modify the external input.

**Context Receptive Generative (CRG) Model:** Our next step was to condition the generative process more concretely on the context  $s$ . We condition the generative process on the representation of  $s$  at each time step. Formally:

$$z_d = \text{Encoder}_d(d_i, \theta_d) \quad (4.4)$$

$$z_s = \text{Encoder}_s(s_i, \theta_s) \quad (4.5)$$

$$\hat{x}_i \sim \prod_t p(\hat{x}_i^t | [\hat{x}_i^{<t}; z_s], z_d) \quad (4.6)$$

where,  $\theta_d$  and  $\theta_s$  are the parameters of the encoder for the document  $d$  and encoder for the context  $s$  respectively,  $z_d$  and  $z_s$  are the encoded representations of the document  $d_i$  and context  $s_i$  respectively. At each time step of generation, the output is conditioned on the tokens generated up to the time step  $t$  concatenated with  $z_s$ . Hence, the generative process is receptive of the context at each time step. This approach uses the Arithmetic or Linear Transform technique to make modifications to the sequential input module (§2.3) of the schema.

### 4.2.2 Extractive models

Generative models that construct new sentences conditioned on the relevant context are compelling but have a number of modeling challenges. Such a model must both select the most relevant content *and* generate a fluent linguistic realization of this information.

We also consider extractive models: approaches that select the most relevant sentence from the document  $d$  to append to the context  $s$ . These approaches can focus solely on the content selection problem and ignore the difficulties of generation. This simplification does come at a cost: the most effective sentence to add might require only a subset of information from some sentence in the document, or incorporate information from more than one sentence.

**Sum-Basic (SB):** One common baseline is Sum-Basic, an extractive summarization technique that relies on word frequency statistics to select salient sentences (Nenkova and Vanderwende, 2005). As an initial step, unigram probabilities are computed from the set of input documents using relative frequency estimation. Then, sentences are selected one-by-one in greedy rounds until the summary budget is saturated. At each round, this model selects the most likely sentence according to the current unigram distribution. The selected sentence is added to the summary and removed from the pool of available sentences. The unigram probabilities of all words in the selected sentence are heuristically discounted (replaced by square root). Select-then-discount operations continue until the summary is written. Discounting is crucial to prevent repetition: once a word (or ideally a concept) has been selected for the summary, it is much less likely to be picked in a subsequent round.

We use Sum-Basic as a Context Agnostic extractive model: we provide the document  $d$  as an input to the model and run Sum-Basic for exactly one round. The selected sentence is considered to be the update sentence  $x$ .

**Context Informed Sum-Basic (CISB):** We developed a simple modification of the Sum-basic technique to incorporate information from the context  $s$  as context. Initial unigram probabilities are computed using word counts from *both* the context *and* the document. Next, for each sentence in the context, we apply just the discount procedure, updating the probability distribution as if those sentences were selected. Finally, we select the single sentence from the document that is most likely according to the resulting discounted unigram probabilities. This simple modification of Sum-Basic helps select a sentence that is novel with respect to the context by lowering the probability of all words already present.

**Extractive CAG, CIG, CRG Models:** Any generative model of  $x$  can also be used as an extractive model: We simply estimate the likelihood of each sentence in the document according to the model, and select the most likely one. Generative models may fail because either they are unable to select the most relevant information, or because the resulting sentence is

ill-formed. Extractive ranking circumvents all errors due to generation and can help isolate model issues.

**Hybrid CAG, CIG, CRG Models:** Since the document  $d$  can be quite large, a generative model may struggle to pick the most salient information based on the context. To simplify the generative modeling task, we can pre-filter the document toward only the most salient parts. We use the Context Informed Sum-Basic technique to first select the top five sentences from the document. We supply only these five sentences in place of the source document  $d$ , then apply the CAG, CIG, and CRG techniques described above.

### 4.2.3 Pre-trained Encoder-Decoder Models

We discuss two ways of building effective representations for pre-trained encoder-decoder models to focus on  $d_i$ : (1) combine encoder representations of  $s_i$  and  $d_i$ , (2) include an additional attention multi-head at each layer of the transformer to specifically focus on the content in  $d_i$ .

Zhao et al. (2020a) introduce the state-of-the-art model for document grounded dialogue generation. As described in (§4.1.1), the chat history serves as the context  $s_i$  and  $x_i$  is the response to be generated. Zhao et al. (2020a) pre-train their architecture on the dialogue specific Reddit (Dziri et al., 2018) dataset and learn separate parameters for encoding  $s_i$  and  $d_i$ . Instead, we employ the recent success of the pre-trained encoder-decoder models (Lewis et al., 2019; Raffel et al., 2019) by using BART (Lewis et al., 2019). One key component of solving this task is to build a representation of the content in the document/s  $d_i$  that is *not* present in the context  $s_i$ . We want to leverage the *SelfAttention* feature of transformers (Vaswani et al., 2017) to build such a representation. Zhao et al. (2020a) further has three components—context processor, knowledge processor and the language model, each of which build distributions over the vocabulary space. A decoding manager is then trained to generate a token based on these three distributions. Since, we use a pre-trained language model as our baseline architecture, we don't use a separate language model component. Instead, we direct our efforts to focus on effectively combining  $s_i$  and  $d_i$ .

**Baseline:** The most straightforward way of using BART for modeling  $p_\theta(x_i | s_i, d_i)$  is to concatenate the tokens of the context  $s_i$  and the document  $d_i$  and pass the concatenated sequence ( $[s_i; d_i]$ ) to the BART encoder, and then the decoder generates  $x_i$ . This is the BART baseline; it already has the advantage of the highly contextualized representations of  $c_i$  and  $d_i$  in comparison with Zhao et al. (2020a). However, fully relying on the self-attention mechanism over the concatenated text would lack the explicit distinction between  $s_i$  and  $d_i$ .

Below, we describe two techniques to efficiently build document focused representations. In Figure 4.3, the method which adds an additional *CrossAttention* multi-head sub-layer to each

layer of the transformer is shown. This attention multi-head specifically focuses on the document  $\mathbf{d}_i$ .

**Context Driven Representation:** One of the sub-task of document grounded generation is to build representation of the content in the document which is not present in the context. We leverage self-attention mechanism to build such a representation. We propose to use two encoder representations for  $\mathbf{s}_i$  and  $\mathbf{d}_i$ . We first define  $\mathbf{h}_d = \text{Encoder}([\mathbf{s}_i; \mathbf{d}_i])$  to get a contextualized representation of  $\mathbf{d}_i$ , conditioning on the context  $\mathbf{s}_i$ .  $\mathbf{h}_d$  is equivalent to the representation used in the BART baseline. We would like representation  $\mathbf{h}_d$  to capture information in the document  $\mathbf{d}_i$  which is not present in the context  $\mathbf{c}_i$ . We then apply the same BART encoder to the context alone:  $\mathbf{h}_s = \text{Encoder}(\mathbf{s}_i)$ . We finally concatenate the encoder outputs  $\mathbf{h} = [\mathbf{h}_s; \mathbf{h}_d]$  before passing them to the BART decoder. This  $\mathbf{h}$  is **Context Driven Representation (CoDR)**. Hence, the decoder gets access to the context representation  $\mathbf{h}_s$  and a representation of the document  $\mathbf{h}_d$ . This method does not require any model architectural modification, and instead the encoder and decoder are fined-tuned to use the multiple input representations.

**Document Headed Attention:** In this section, we describe **Document Headed Attention (DoHA)** to further enhance the use of the multiple input representations. A decoder in transformer encoder-decoder models (Vaswani et al., 2017) has two types of multi-head attention mechanism, *SelfAttention* and *CrossAttention* with the source sequence. *SelfAttention* module allows each position in the decoder to attend to all positions in the decoder up to and including that position. *CrossAttention* module performs multi-head attention over the output of the encoder stack and attends over the source sequence. While our CoDR method uses the two different source representations,  $\mathbf{h}_s$  and  $\mathbf{h}_d$ , *CrossAttention* is still shared over the concatenated representation  $\mathbf{h}$ .

In this work, we add an additional multi-head attention *CrossAttention\_Doc* to specifically attend over the tokens of the document, while the original *CrossAttention* (named as *CrossAttention\_Cxt*), only attends over the tokens of the context.

Each of the multi-heads are of the form:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= [\mathbf{H}_1; \dots; \mathbf{H}_m] \mathbf{W}^o, \\ \mathbf{H}_j &= \text{Attention}(Q \mathbf{W}_j^Q, K \mathbf{W}_j^K, V \mathbf{W}_j^V). \end{aligned}$$

The multi-head function receives three inputs - a query  $Q$ , key  $K$  and value  $V$ .  $\mathbf{W}^o$  is an output projection of the concatenated outputs of the attention heads. Each  $\mathbf{H}_j$  is the output of a single attention head and  $\mathbf{W}_j^Q$ ,  $\mathbf{W}_j^K$  and  $\mathbf{W}_j^V$  are head-specific projections for  $Q$ ,  $K$ , and  $V$ , respectively.



Hence, the multi-head *CrossAttention.Doc* is defined by:

$$\begin{aligned} \text{CrossAttention.Doc}(Q, K, V) &= [\mathbf{H}_1; \dots; \mathbf{H}_m] \mathbf{W}^{\text{do}}, \\ \mathbf{H}_j &= \text{Attention}(Q \mathbf{W}_j^{\text{dQ}}, K \mathbf{W}_j^{\text{dK}}, V \mathbf{W}_j^{\text{dV}}), \end{aligned}$$

where  $\mathbf{W}^{\text{do}}$ ,  $\mathbf{W}_j^{\text{dQ}}$ ,  $\mathbf{W}_j^{\text{dK}}$  and  $\mathbf{W}_j^{\text{dV}}$  are parameters trained specifically to focus on document. The parameters of *CrossAttention.Doc* are initialized with those of *CrossAttention.Cxt*.

Each decoder layer follows the following sequence of functions:

$$\begin{aligned} \mathbf{h} &= \mathcal{F}(\text{SelfAttention}(\mathbf{h}_x, \mathbf{h}_x, \mathbf{h}_x)), \\ \mathbf{h} &= \mathcal{F}(\text{CrossAttention.Cxt}(\mathbf{h}, \mathbf{h}_s, \mathbf{h}_s)), \\ \mathbf{h} &= \mathcal{F}(\text{CrossAttention.Doc}(\mathbf{h}, \mathbf{h}_d, \mathbf{h}_d)), \\ \mathbf{h} &= \mathcal{F}(\text{FFN}(\mathbf{h})), \end{aligned}$$

where  $\mathcal{F}(\mathbf{h})$  is a sequence of  $\text{LayerNorm}(\text{residual} + \text{dropout}(\mathbf{h}))$ , followed by  $\text{residual} = \mathbf{h}$ . We integrate the additional attention head *CrossAttention.Doc* by passing the output of the previous attention head *CrossAttention.Cxt* as query. Unlike the weighted attention fusion techniques (Cao et al., 2020), this technique of fusing the additional attention head is novel and useful as it does not require any additional parameters for the fusion.

## 4.3 Experiments

We evaluate the models using both automated metrics and, for a subset of promising systems, human assessment. One key evaluation is the similarity between the model generated sentence and reference sentence. Another crucial evaluation is the notion of grounding in the document. We evaluate if the generated sentence is coherent to the context and contains information from the document using human evaluation. Human judges are also asked to assess grammaticality and coherence.<sup>6</sup>

### 4.3.1 Automated Evaluation

The primary automated evaluation metric for document grounded generation is ROUGE-L F1 against reference sentence, though we also include BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2011) as additional indicators. ROUGE is a standard family of metrics for summarization tasks; ROUGE-L measures the longest common subsequence between the system and the reference, capturing both lexical selection and word order. METEOR also uses synonyms and stemmed forms of the words in candidate and reference sentences,

<sup>6</sup>Details about hyper-parameters, generated examples and examples of human dialogues are provided in Appendix B.

Model	ROUGE-L	BLEU	METEOR
Sum-Basic	5.6 (5.6–5.7)	0.6	2.0
Context Informed Sum-Basic (CISB)	7.0 (7.0–7.1)	1.0	2.8
Context Agnostic Generative (CAG)	9.1 (9.0–9.2)	1.2	4.6
Context Only Generative (COG)	13.5 (13.4–13.6)	1.7	3.5
Context Informed Generative (CIG)	<b>16.0</b> (15.9–16.1)	<b>3.5</b>	<b>5.3</b>
Context Receptive Generative (CRG)	14.7 (14.6–14.8)	2.6	4.5
Hybrid CAG	8.0 (7.9–8.0)	1.0	3.8
Hybrid CIG	15.0 (14.9–15.1)	2.7	4.7
Hybrid CRG	13.5 (13.4–13.6)	2.3	4.1
Extractive CAG	9.3 (9.2–9.3)	1.1	3.2
Extractive CIG	9.3 (9.2–9.3)	1.1	3.2
Extractive CRG	9.2 (9.1–9.3)	1.1	3.2
<i>Oracle</i>	28.8 (28.7–29.0)	11.0	10.9

TABLE 4.7: Results on automated metrics for Wikipedia Update Generation task; 95% confidence interval in parentheses.

and thus may be better at quantifying semantic similarities. Additionally, we present F1 which indicates the unigram overlap between the generated output and the reference sentence.<sup>7</sup>

Table 4.7 presents results for baseline models for the Wikipedia Update Generation task.<sup>8</sup> It illustrates that this task is quite difficult for extractive techniques. Furthermore, the results emphasize the importance of having curated text as context when generating the update. In all experimental conditions, models aware of context perform much better than models agnostic of it. In contrast to Liu et al. (2018), generative approaches outperformed hybrid, likely because we only had a single input document. Extractive CAG, CIG, and CRG all outperformed both Sum-Basic and the context informed variant. Extractive CAG was on-par with generative CAG, suggesting the generated sentences were of reasonable quality. However, generative CIG and CRG were substantially better: rewriting to match context was beneficial.

The *Oracle* system of Table 4.7 aims to establish an upper limit attainable by extractive methods, using the following oracle experiment: For each test instance  $(d_i, s_i, x_i)$ , we enumerate each extracted sentence  $e$  of document  $d_i$  and select the one with highest ROUGE-L score as *Oracle*'s update sentence  $\hat{x}_i$  (i.e.,  $\hat{x}_i = \arg \max_{e \in d_i} \text{ROUGE-L}(x_i, e)$ ).

Note this yields a very optimistic upper bound, as the same ground truth  $x_i$  is used both to select an extractive sentence from a large pool of candidates and for final automatic metric scoring.<sup>9</sup> Nevertheless, these oracle results let me draw two conclusions: (1) They give a

<sup>7</sup>We use the code published at <https://github.com/facebookresearch/ParlAI/blob/master/parlai/core/metrics.py> to calculate unigram F1.

<sup>8</sup>We use the pyrouge toolkit along with ROUGE-1.5.5: <https://github.com/bheinzerling/pyrouge>

<sup>9</sup>Previous work has shown that this type of oracle can yield upper bounds that are unrealistically high, and they tend to be above human performance (Och et al., 2004, Table 1). One remedy suggested by Och et al. (2004) is a round-robin oracle ensuring that the reference (ground truth) used by the argmax is distinct from that of the final automatic evaluation, but that scheme is only possible with a multi-reference test set.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	Meteor	F1
<b>Wikipedia Update Generation</b>							
CIG	10.18	4.42	2.20	1.23	10.08	6.21	12.6
BART (baseline)	21.72	14.71	11.28	9.20	22.39	12.90	27.5
CoDR	<b>25.15</b>	<b>17.33</b>	<b>13.56</b>	<b>11.31</b>	23.48	<b>14.38</b>	29.0
DoHA	25.11	17.04	13.17	10.86	<b>23.49</b>	14.28	29.1
<b>CMU DoG</b>							
LowR (Zhao et al., 2020a)	15.00	5.70	2.50	1.20	-	-	10.7
BART (baseline)	23.78	19.27	17.66	16.91	19.30	12.59	21.7
CoDR	26.86	22.75	21.30	20.68	20.41	14.47	22.7
DoHA	<b>27.33</b>	<b>23.05</b>	<b>21.55</b>	<b>20.90</b>	<b>20.44</b>	<b>14.55</b>	22.8
<b>Wizard of Wikipedia (Seen)</b>							
LowR (Zhao et al., 2020a)	21.80	11.50	7.50	5.50	-	-	18.0
BART (baseline)	23.92	14.62	10.24	7.75	21.41	15.45	31.1
CoDR	24.00	14.98	10.64	<b>8.18</b>	<b>21.82</b>	15.71	31.8
DoHA	<b>24.14</b>	<b>15.08</b>	<b>10.68</b>	<b>8.18</b>	21.76	<b>15.89</b>	31.8
<b>Wizard of Wikipedia (Unseen)</b>							
LowR (Zhao et al., 2020a)	20.70	10.10	6.20	4.30	-	-	16.5
BART (baseline)	21.88	12.54	8.44	6.23	19.14	14.03	28.2
CoDR	21.84	12.74	8.60	6.35	19.50	14.22	29.0
DoHA	<b>22.31</b>	<b>13.04</b>	<b>8.89</b>	<b>6.60</b>	<b>19.62</b>	<b>14.47</b>	29.0

TABLE 4.8: Results on the automated metrics for the three datasets

better perspective to assess the non-oracle systems, and we believe that their seemingly low automatic evaluation scores are quite reasonable relative to the optimistic upper bound (e.g., CIG’s ROUGE-L’s score is 55% of the oracle). (2) The oracle results suggest that humans are substantially changing the surface realization as they summarize for Wikipedia, as otherwise the oracle results would be much closer to maximum metric scores (i.e., 100%). This shows that extractive methods are not enough for this task, justifying our use of generation techniques.

To automatically evaluate the fluency of the baseline models for Grounded Dialogue Generation task, we use the perplexity measure. We build a language model on the train set of responses using ngrams up to an order of 3<sup>10</sup>. The Context Only Generative (COG) model which generates the dialogue response based on the previous dialogue turn only, achieves a perplexity of 21.8. The Context Receptive Generative (CRG) model on the other hand, which provides the section information as an additional input to each time step of the decoder, achieves a perplexity of **10.11**. This indicates that including the sections of document helps in the generation process.

Table 4.8 presents results for all the three tasks on BART-based models.<sup>11</sup> It shows that the BART baseline outperforms previous state-of-the-art models (Zhao et al., 2020a; Prabhumoye et al., 2019b) on all three tasks. It demonstrates that both our improvements DoHA and CoDR perform better than our BART baseline on all metrics and for all three tasks. Notably, we see an

<sup>10</sup>We use the SRILM toolkit (Stolcke, 2002)

<sup>11</sup>We use NLG evaluation toolkit (Sharma et al., 2017) from <https://github.com/Maluuba/nlg-eval>

Evaluation task	prefer		
	CAG	neither	CIG
Close to reference	15.8%	53.3%	30.8%
Coherent to context	7.5%	53.3%	39.2%

TABLE 4.9: Human preferences of CAG vs. CIG.

improvement of 19.7 BLEU-4 points on the CMU\_DoG dataset compared to Zhao et al. (2020a) which was pre-trained on dialogue specific data; and an improvement on 8.9 BLEU-4 points on the Wikipedia Update Generation compared to (Prabhumoye et al., 2019b).<sup>12</sup> We also see substantial improvements (23.6% increase in BLEU-4 for CMU\_DoG) compared to the simple BART baseline for the three tasks. In general, DoHA performs slightly better than CoDR on the three tasks.

### 4.3.2 Human Evaluations

For careful evaluation of the performance of the most promising configurations, we also asked human judges for quality assessments. We solicited several types of evaluation, including two relative comparisons between pairs of system outputs and an absolute quality evaluation of individual system outputs. We evaluate the system generated sentences on three dimensions: (1) closeness of the generated sentences to the references, (2) relevance of the generated sentences to the context and document, and (3) fluency of the generated sentences.

**Closeness:** The automatic metrics like BLEU, METEOR, and Rouge-L may not be tolerant towards linguistic variations in generated outputs. Hence, we perform a human evaluation to measure how accurately the generated sentence reflects the information in the reference. The annotators are provided with the reference sentence and the generated outputs of two systems labeled *A* and *B* in a randomized order. The annotators were instructed to “Pick the option which is closest in meaning with the reference option.” The annotators could select system *A* or *B*, or indicate that neither was preferred by picking the third option *C*. This is a simple evaluation task though potentially biased toward the sole reference.

**Relevance:** The reference sentence may not be the only correct sentence that fits the context. This is especially true in dialogue generation tasks where contexts like “How are you?” and “What was your favourite part of the movie?” can have many correct responses that can be produced by grounding on the same document. Hence, we measure whether the generated output contained salient information from the document written in a manner appropriate to the context. The annotators are provided with the document  $\mathbf{d}_i$ , the context  $\mathbf{c}_i$ , and the outputs of the two systems *A* and *B*, again in a random order. They were instructed to “Pick the option

<sup>12</sup>We use NLG eval script for (Prabhumoye et al., 2019b)

Task	BART v CoDR			BART v DoHA			DoHA v CoDR		
	BART	NoPref	CoDR	BART	NoPref	DoHA	DoHA	NoPref	CoDR
<b>Wikipedia Update Generation</b>									
<i>Closeness</i>	33.3	36.7	30.0	25.5	46.7	27.8	32.2	42.2	25.6
<i>Relevance</i>	18.9	54.4	26.7	24.4	45.6	30.0	33.3	38.9	27.8
<b>CMU_DoG</b>									
<i>Closeness</i>	15.6	58.8	25.6	30.0	42.2	27.8	33.3	44.5	22.2
<i>Relevance</i>	22.2	43.4	34.4	23.3	42.3	34.4	34.4	42.3	23.3
<b>Wizard of Wikipedia (seen)</b>									
<i>Closeness</i>	36.7	40.0	23.3	28.9	31.1	40.0	40.5	31.7	27.8
<i>Relevance</i>	24.2	51.6	24.2	32.2	35.6	32.2	28.9	46.7	24.4
<b>Wizard of Wikipedia (unseen)</b>									
<i>Closeness</i>	23.3	47.8	28.9	44.4	20.0	35.6	21.1	63.3	15.6
<i>Relevance</i>	27.8	47.8	24.4	30.0	43.3	26.6	23.3	41.1	35.6

TABLE 4.10: Human evaluation results depicting percentage of times a model was picked (No-Pref=No Preference)

which contains information from the document and fits the dialogue context coherently”. Note that the annotators don’t have access to the reference in this evaluation. Each judge had to consider whether the information fits with the context and also whether system-generated content could be supported by the document.

Table 4.9 shows the results for baseline models on the Wikipedia Update Generation task: the context-aware CIG system was substantially better in both settings. For these result, four human judges each annotated 30 unique output pairs for these two relative comparison settings, a total of 240 relative judgments.

Table 4.10 shows the results of the human evaluation on closeness and relevance for all the three tasks for the BART-based models. This human evaluation was conducted on Amazon Mechanical Turk. We conduct 3 comparative studies between the BART, CoDR and DoHA outputs. Each worker was asked to annotated 10 pairs of sentences. We added one control pair among them i.e for 1/10 pairs, both the sentences were exactly the same. If a worker provides wrong judgement for the control pair then their annotations were discarded. For each dataset we have total 540 comparative judgements and 90 sentences of each of the models marked for fluency.

The closeness results show that all the three models BART, CoDR and DoHA generate sentences that are close to the reference, although CoDR and DoHA outperform BART in most cases. Interestingly, the relevance results for Wikipedia Update Generation and CMU\_DoG datasets show that CoDR and DoHA generate content that is grounded in the document as opposed to BART. BART baseline generates sentences that are fluent and close to the reference but does not ground in the content of the document as compared to CoDR and DoHA. The ‘No Preference’ is generally opted over any of the models which is further discussed in §4.3.3. For the relevance comparison, annotators have to read a large document to figure out if the

Model	Grammaticality	Non-redundancy	Referential Clarity	Focus	Structure
CAG	2.6	1.8	2.7	2.6	2.4
CIG	<b>4.3</b>	<b>3.9</b>	<b>3.6</b>	<b>3.5</b>	<b>3.2</b>

TABLE 4.11: Human absolute quality assessments.

generated information is present in the document or not. This can make the annotations noisy especially for Wizard of Wikipedia dataset which has 7 passages as grounding document.

**CoDR and DoHA:** The DoHA model still uses the content driven representations ( $\mathbf{h}_d$  and  $\mathbf{h}_s$ ). The main difference is that in CoDR model we concatenate  $\mathbf{h}_d$  and  $\mathbf{h}_s$  and pass it to the decoder but for DoHA we pass  $\mathbf{h}_d$  and  $\mathbf{h}_s$  separately to the decoder. DoHA has an additional MHA layer to focus on the representation of the document  $\mathbf{h}_d$  only. In this loose sense, DoHA is CoDR plus additional parameters in MHA to focus on  $\mathbf{h}_d$ . DoHA performs marginally better than CoDR in automated metrics. But qualitatively (human evaluation) DoHA produces higher quality outputs as compared to CoDR. Table 4.10 shows DoHA performing better than CoDR on all but one case.

**DUC Guidelines (Absolute):** In addition, we performed an absolute quality evaluation following the guidelines from DUC 2007 for the baseline models for the Wikipedia Update Generation task.<sup>13</sup> Each judge was presented with a single system output, then they were asked to evaluate five aspects of system output: grammaticality, non-redundancy, referential clarity, focus, and structure/coherence. For each aspect, the judge provided an assessment on a five-point scale: (1) Very Poor, (2) Poor, (3) Barely Acceptable, (4) Good, (5) Very Good. We gathered 120 additional judgments in this setting (4 judges, 30 outputs). Again, context-aware CIG substantially outperforms CAG across the board, as seen in Table 4.11.

**Fluency:** Finally, we evaluate the fluency of the generated sentences on a scale of 1 (unreadable) to 4 (perfect) as is described in (Zhou et al., 2018). For the baselines of the Grounded Dialogue Generation task, we randomly select 120 generated responses from each model and each response was annotated by 3 unique workers. The COG model got a low score of 2.88, in contrast to the CRG score of **3.84**. This outcome demonstrates that the information in the section also helps in guiding the generator to produce fluent responses. Since both CoDR and DoHA are also BART-based models, the fluency for all three of them is very high and close to each other (BART=3.64, CoDR=3.71, DoHA=3.66). The BART-based results are aggregated across all the three tasks.

**Engagement:** In addition to the above metrics, we measure engagement of the generated response for the Grounded Dialogue Generation task. We set up a pairwise comparison following

<sup>13</sup><http://duc.nist.gov/duc2007/quality-questions.txt>

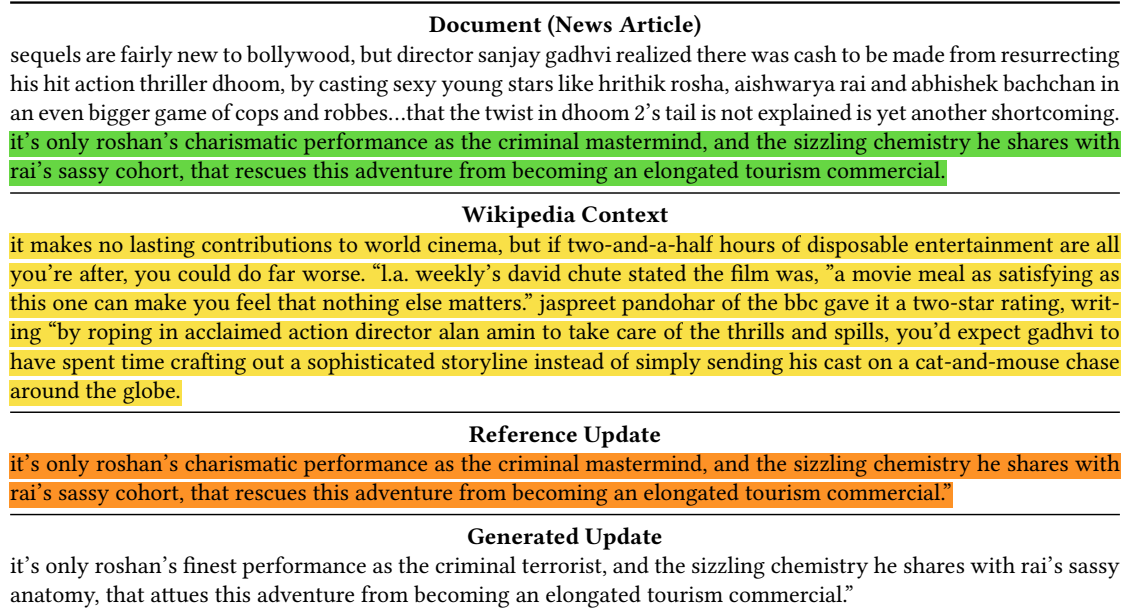


FIGURE 4.5: Example of good quality generation, where the system-generated update is close to the reference.

Bennett (2005) to evaluate the engagement of the generated responses. The test presents the chat history (1 utterance) and then, in random order, its corresponding response produced by the COG and CRG models. A third option "No Preference" was given to participants to mark no preference for either of the generated responses. The instruction given to the participants is "Given the above chat history as context, you have to pick the one which can be best used as the response based on the engagingness." We randomly sample 90 responses from each of the COG and CRG models. Each response was annotated by 3 unique workers and we take majority vote as the final label. The result of the test is that COG generated responses were chosen only 36.4% times as opposed to CRG generated responses which were chosen 43.9% and the "No Preference" option was chosen 19.6% of times. This result shows the information from the sections improves the engagement of the generated responses.

### 4.3.3 Manual Inspection

**Preliminary observations on the Wikipedia Update Generation task:** Systems unaware of the context  $s$  tend to generate long updates with repeated frequent words or phrases. Consider the ratio of unique tokens over the total number of tokens in the generated output, which is denoted by  $R$ . A small  $R$  indicates many repeated tokens. We find that 88% of the time this ratio  $R$  falls below 0.5 for the CAG model, i.e. for 88% instances, more than 50% of the words in the generated output are repeats. This number is relatively small – 14% for CIG and 20% for CRG – in context aware models. In the reference updates only 0.21% instances repeat more than 50% of words.

Document (News Article)
anne kirkbride, who portrayed bespectacled, gravelly-voiced deirdre barlow in coronation street for more that four decades, <b>has died. the 60-year-old</b> , whose first appearance in the soap opera was in 1972, died in <b>a manchester hospital</b> after a short illness.... kirkbride had left the soap opera after <b>she was diagnosed with non-hodgkin's lymphoma</b> in 1993 but returned some months later after treatment and spoke candidly about how she had struggled with depression following the diagnosis...
Wikipedia Context
in 1993, kirkbride was diagnosis with non-hodgkin's lymphoma. she spoke to the british press about her bout of depression following the diagnosis. she was cured within a year of being diagnosed.
Reference Update
anne kirkbride died of breast cancer in a manchester hospital on 19 january 2015, aged 60.
Generated Update
she was diagnosed with non-hodgkin's lymphoma.

FIGURE 4.6: Example of lower-quality output: the generated update unnecessarily restates information yet misses the most salient detail from the document.

Reference Update	Generated Update
1. rob brydon, the comedian was born in baglan.	he was born in baglan.
2. in may 2014 he was diagnosed with prostate cancer.	st. clair was diagnosed with prostate cancer.
3. he was the first black player to have played a game in the national basketball association.	he was the first african-american to play in the national basketball association.
3. on april 3, 2014, manning signed a one-year deal with the cincinnati bengals.	on march 9, 2014, manning signed a one-year contract with the cincinnati bengals.
4. on oct 10, 2013, barrett signed with the memphis grizzlies.	on feb 9, 2013, barrett signed with the memphis grizzlies.
6. anne kirkbride died of breast cancer in a manchester hospital on 19 january 2015, aged 60.	she was diagnosed with non-hodgkin's lymphoma.
5. some people think elvis is still alive, but most of us think he's dead and gone."	some people think elvis, but most of us think he's dead and gone."
6. it's always the goal of the foreign-language film award executive committee to be as inclusive as possible."	it's always the goal of the foreign- entry film award executive to be as possible."

TABLE 4.12: Example generations from the CIG system, paired with the human generated updates.

Figures 4.5 and 4.6 show good and bad examples generated by the CIG model along with the document, context and the reference update. Table 4.12 has a set of updates generated by the CIG model as well as the reference update. As can be seen in examples 3 and 4, the CIG model misplaces the date but correctly generates the remaining content. In examples 1 and 2, the CIG model appears to successfully select the correct pronouns for co-reference resolution, though it gets confused as to when to use the pronoun or the named entity. Examples 5 and 6 represent failure cases due to missing words.



Error Class	%	Chat context	Reference	Generation
<i>Reference and generation are grounded</i>	35	the story is sounding even more interesting. who do you think had the most growth or development in the film?	but there's a twist in that <b>anna has been deceived by the man she was in love with, he was really plotting to take over the kingdom from the sisters the whole time.</b>	but there's a twist in that <b>anna is actually in love with a man who is actually a bad guy. he was actually trying to take over the kingdom from the sisters after elsa injured anna and then their parents died.</b>
<i>Reference is grounded but generation is not</i>	3	hello!	have you heard about <b>the post directed by spielberg?</b>	hello!
<i>Reference is not grounded but generation is</i>	5	yes i agree and how meryl can depict the first female publisher. that would be interesting to see the dynamic of how it effected her career as well	yes, for a woman to have a successful career in those times was not an easy task, for sure.	definitely. i wonder how <b>mcnamara</b> felt when he saw the documents and then <b>told ellsberg that the war was hopeless.</b>
<i>Reference and generation are not grounded</i>	43	the shark was so realistic	yea i live near a beach and it made me not ever want to go swimming	i never thought of it that way, and i agree, it made me terrified to go into the ocean for awhile.

TABLE 4.13: Error analysis for the CMU\_DoG dataset (only shown for outputs with low Rouge-L score). Text in green indicates grounded information.

**Error analysis on the BART-based models:** We manually inspect the outputs of the CoDR model on the development set of CMU\_DoG and Wikipedia Update Generation dataset to understand the their quality. We inspect 60 samples in each dataset which have Rouge-L score  $< 60$ . These are chosen such that we have 10 samples in each of the 6 buckets of Rouge-L score (buckets are range of 10 points: 0-9, 10-19, 20-29, 30-39, 40-49 and 50-59). We analyse the generated outputs along the two aspects of appropriateness of the generation to the context and its grounding in the document.

**CMU\_DoG:** We find that 52/60 (86.7%) responses were appropriate to the given chat context. These 52 responses are further categorized in Table 4.13. We found that for about 90% of samples, if the reference is grounded then the generation is also grounded and if the reference is not grounded then the generation is not grounded. Further inspection shows that references are not grounded if they are follow up questions, opinions or experiences that are shared in the conversation. In most of these cases, the context dictates if the response should be grounded or not grounded in the document. Since, all of the generated responses in this category are appropriate to the context suggests that these conversational subtleties are not captured by automated evaluation metrics and are given a low score. We also observe a few data artifacts like the mapping of the Wikipedia sections and the chat context is noisy for this dataset. This

Error Class	%	Reference	Generation	R
<i>Linguistic Variation:</i> Reference and generation are grounded and generation is appropriate but a linguistic variation of the reference or an alternate appropriate update.	43	December 12 - The Smiths play Brixton Academy, their last ever gig before their dissolution.	December 12 - The Smiths perform their final show, at Brixton Academy in London.	41
<i>Partial Hallucination:</i> Reference and generation are grounded but generation is either missing or hallucinates some information	23	America Online and Prodigy (online service) offered access to the World Wide Web system for the first time this year, releasing browsers that made it easily accessible to the general public.	The World Wide Web was first introduced on January 17, 1995 on Prodigy.	17
<i>Incoherent Reference:</i> The reference does not coherently follow the context	22	“The Naked Ape”, by Desmond Morris, is published.	Zoologist Desmond Morris publishes “The Naked Ape”.	26
<i>Incorrect:</i> The generation is either not appropriate or is not grounded (completely hallucinates the information).	7	The year 2000 is sometimes abbreviated as “Y2K” (the “Y” stands for “year”, and the “K” stands for “kilo-” which means “thousand”).	The Y2K conspiracy theory claimed that a secret nuclear attack by the United States on 2 January 2000 was planned to begin World War 2.	9
<i>Reference is not grounded</i>	5	This was achieved under dead calm conditions as an additional safety measure, whereas the Wrights flew in a 25 mph+ wind to achieve enough airspeed on their early attempts.	This was verified by a video crew present at the test flight.	14

TABLE 4.14: Error Analysis for Wikipedia Update Generation task (R denotes Rouge-L score. Text in red indicates hallucinated or missing information.)

can be easily resolved by providing all the previous passages of the conversation as grounding to the model. We would also like to note that this dataset was collected under two scenarios: (1) both the people in the conversation have access to the document, and (2) only one person has access to the document. But this distinction is not made in modeling the task. The noise in the dataset can be reduced by modeling only the users that have access to the document in the conversation (similar to Wizard of Wikipedia where only the wizard is modeled).

**Wikipedia Update Generation:** The error analysis for this task is shown in Table 4.14. For 5% cases, the reference itself is not grounded in the document. The remaining 95% cases are further classified into 4 error categories. About 85% times, the generation is either completely or partially grounded if the reference is grounded. 43% generations are grounded in document but are linguistic variations of the reference or could be alternate updates to the context. Yet, these are scored low on the Rouge-L metric revealing the inadequacy of the automated metrics.

Dataset	Reference	No Preference	DoHA and/or CoDR
Wikipedia Update Generation	33.9	28.3	37.8
CMU_DoG	22.8	45.6	31.6

TABLE 4.15: Comparison with reference (Ref) in %age

For 23% cases the generation partially hallucinates some information or misses some information present in the reference. 22% times the reference itself does not seem to coherently fit the context. This is primarily observed for Wikipedia pages that are in the form of a list like *1340s* and *Timeline of DC Comics (1950s)*. Yet, for 50% of the *Incoherent Reference* cases, the generation is grounded in the document and very close to the reference (like the example in Table 4.14). Only for 7% of the cases, the generation is completely incorrect and hallucinates all of the information. Future work can focus on improving the error in the *Incorrect* and *Partial Hallucination* error classes.

**Reference Comparison:** With the insights from manual inspection, we performed another comparative study with human judges (on Amazon Mechanical Turk). This was to understand how close our models perform in comparison with the reference. The judges are instructed to “Pick the option that is most appropriate to the given context”. We annotated 100 samples for each DoHA and CoDR model in comparison with the reference on the CMU\_DoG and Wikipedia Update Generation datasets. The results in Table 4.15 show consolidated results for the two models. It demonstrates that our models produce appropriate outputs which can be used as alternate updates/responses. Our models are preferred over the reference in both the tasks suggesting that the automated evaluation is insufficient and the sole reference should not be considered as the only correct response to the context.

## 4.4 Ethical Considerations

The intended use of the models proposed is to aid the NLG systems in generating content-rich text. Note that this does not imply that the models generate factually correct text. The generation entirely depends on the information in the document provided. If the document itself is factually incorrect then the generation would be grounded in false content and hence generate inaccurate text.

We hope that this technology is used for socially positive applications like building trust of users in dialogue systems like Alexa, Siri and Google Home by providing users with credible information. This chapter has specifically focused on dialogue response generation task with the aim that this research not only helps in generating responses which contain useful information but also increase credibility of responses by disclosing the source of information. If dialogue systems base their responses on certain sources of information then they can potentially disclose the source of information to the user. The user then has the agency to make informed decision about trusting the system responses or not.

Document	Context	Generation
<a href="https://www.bradenton.com/latest-news/article237480339.html">https://www.bradenton.com/latest-news/article237480339.html</a>	Climate change includes both the global warming driven by human emissions of greenhouse gases, and the resulting large-scale shifts in weather patterns. Though there have been previous periods of climatic change, since the mid-20th century the rate of human impact on Earth's climate system and the global scale of that impact have been unprecedented.	Though most scientists agree that a "global" climate change is occurring, climate change remains a contentious topic and many scientists, climate scientists, and even climate skeptics believe that it is a hoax.
<a href="https://tinyurl.com/3r5runb">https://tinyurl.com/3r5runb</a>	Climate change includes both the global warming driven by human emissions of greenhouse gases, and the resulting large-scale shifts in weather patterns. Though there have been previous periods of climatic change, since the mid-20th century the rate of human impact on Earth's climate system and the global scale of that impact have been unprecedented.	Some scientists dismiss the existence of climate change.
<a href="https://tinyurl.com/y634czdp">https://tinyurl.com/y634czdp</a>	Climate change includes both the global warming driven by human emissions of greenhouse gases, and the resulting large-scale shifts in weather patterns. Though there have been previous periods of climatic change, since the mid-20th century the rate of human impact on Earth's climate system and the global scale of that impact have been unprecedented.	The Scientific opinion on climate change is that planetwide warming has occurred at least three million years ago.

TABLE 4.16: Outputs generated by DoHA model for a few hand selected news articles. The DoHA model is trained on Wikipedia Update Generation dataset.

Tables 4.16 and 4.17 shows the potential misuses of models trained on this task. For both the experiments, a few news articles were hand selected and relevant context was selected from a chosen Wikipedia article. In case of 4.17, the context was curated by hand. The inference was generated using trained models in each domain.

Interestingly, Table 4.16 also shows the sensitivity of the trained model to the document information. It consists of the same context but different documents were provided as inputs to the model. The generated outputs are different for each new document.

Table 4.17 shows examples generated by DoHA model trained on Wizard of Wikipedia dataset. This model was chosen as opposed to the model trained on CMU\_DoG dataset because the CMU\_DoG dataset only consists of dialogues on the topic of movies. This table too shows the sensitivity of the model towards the document and that the model relies on the information in the document for generation.

Document	Context	Generation
<a href="https://tinyurl.com/yxedlapz">https://tinyurl.com/yxedlapz</a>	What do you think of the rising sea levels?	I think it's crazy that people believe that the oceans will rise and yet people are investing \$15 million to live on islands...
<a href="https://tinyurl.com/y3u6fhpo">https://tinyurl.com/y3u6fhpo</a>	Do you think the Affordable Care Act is helping the general public?	It has some great benefits. It has actually resulted in millions of Americans losing their health insurance plan. That is not something former VP Biden claimed.
<a href="https://tinyurl.com/3r5runb">https://tinyurl.com/3r5runb</a>	What do you think of the rising sea levels?	I think the rise of sea levels is a big money making hoax. Why do you think the world is heating up? Any experts?

TABLE 4.17: Outputs generated by DoHA model for a few hand selected news articles. The DoHA model is trained on Wizard of Wikipedia dataset.

## 4.5 Related Work

### 4.5.1 Task

**Wikipedia Update Generation:** The proposed content transfer task is clearly related to a long series of papers in summarization, including recent work with neural techniques (Rush et al., 2015; Nallapati et al., 2016). In particular, one recent paper casts the the task of generating an entire Wikipedia article as a multi-document summarization problem (Liu et al., 2018). Their best-performing configuration was a two-stage extractive-abstractive framework; a multi-stage approach helped circumvent the difficulties of purely abstractive methods given quite large input token sequences.

Looking beyond the clear task similarity of authoring Wikipedia style content, there are several crucial differences in our approach. First, the goal of that paper is to author the whole page, starting from nothing more than a set of primary sources, such as news articles. In practice, however, Wikipedia articles often contain information outside these primary sources, including common sense knowledge, framing statements to set the article in context, and inferences made from those primary sources. Our task restricts the focus to content where a human editor explicitly decided to cite some external source. Hence, it is much more likely that the resulting summary can be derived from the external source content. Furthermore, we focus on the act of adding information to existing articles, rather than writing a complete article without any context. These two scenarios are clearly useful yet complementary: sometimes people want to produce a new reference text where nothing existed before; in other cases the goal is to maintain and update an existing reference.

Another closely related task is update summarization (Dang and Owczarzak, 2008), where systems attempt to provide a brief summary of the novel information in a new article assuming the user has read a known set of prior documents. Our focus on curating an authoritative

resource is a substantial difference. Also our datasets are substantially larger, enabling generative models to be used in this space, where prior update summarization techniques have been primarily extractive (Fisher and Roark, 2008; Li et al., 2015).

For any generation task, it is important to address both the content (‘what’ is being said) as well its style (‘how’ it is being said). Recently, a great deal of research has focused on the ‘how’ (Li et al., 2018a; Shen et al., 2017), including efforts to collect a parallel dataset that differs in formality (Rao and Tetreault, 2018), to control author characteristics in the generated sentences (Prabhumoye et al., 2018), to control the perceived personality traits of dialog responses (Zhang et al., 2018). We believe this research thread is complementary to our efforts on generating the ‘what’.

Another form of content transfer bridges across modalities: text generation given schematized or semi-structured information. Recent research has addressed neural natural language generation techniques given a range of structured sources: selecting relevant database records and generating natural language descriptions of them (Mei et al., 2016), selecting and describing slot-value pairs for task-specific dialog response generation (Wen et al., 2015), and even generating Wikipedia biography abstracts given Infobox information (Lebret et al., 2016). Our task, while grounded in external content, is different in that it leverages *linguistic* grounding as well as prior text context when generating text. This challenging setting enables a huge range of grounded generation tasks: there are vast amounts of unstructured textual data.

**Document Grounded Dialogue Response Generation:** Dialog systems are considered to be either task-oriented, where a specific task is the goal of the conversation (e.g. getting bus information or weather for a particular location); or non-task oriented where conversations are more for the sake of themselves, be it entertainment or passing the time. Ultimately, we want our agents to smoothly interleave between task-related information flow and casual chat for the given situation. There is a dire need of a dataset which caters to both these objectives.

Serban et al. (2015) provide a comprehensive list of available datasets for building end-to-end conversational agents. Datasets based on movie scripts (Lison and Tiedemann, 2016; Danescu-Niculescu-Mizil and Lee, 2011) contain artificial conversations. The Ubuntu Dialogue Corpus (Lowe et al., 2015) is based on technical support logs from the Ubuntu forum. The Frames dataset (El Asri et al., 2017) was collected to solve the problem of frame tracking. These datasets do not provide grounding of the information presented in the conversations. Zhang et al. (2018) focuses on personas in dialogues: each worker has a set of predefined facts about the persona that they can talk about. Most of these datasets lack conversations with large number of on-topic turns. We introduce a new dataset which addresses the concerns of grounding in conversation responses, context and coherence in responses. We present a dataset which has real human conversations with grounding in a document. Although the examples use Wikipedia articles about movies, the same techniques being valid for other external documents such as manuals, instruction booklets, and other informational documents. Dinan et al. (2018) also

introduce a dataset of human-human conversation that are grounded in Wikipedia articles. These conversations are grounded in a diverse range of topics (totally 1365).

### 4.5.2 Methodology

Generation grounded in document has been studied through a large body of summarization work (Rush et al., 2015; Nallapati et al., 2016) and similar tasks such as headline generation (Tan et al., 2017). Multiple new works have extended this research in new directions; Wikipedia Update Generation (Prabhumoye et al., 2019b) introduces the task of generating an *update* to the Wikipedia context based on a news document; Wikipedia article generation (Liu et al., 2018) introduces the task of generating an entire Wikipedia article based on multiple documents; Text Editing by Command (Faltings et al., 2020) introduces the task of generating a particular type of Wikipedia edit conditioned on a command provided in natural language and a grounding consisting of snippets of 200 web page results.

Parallely, new tasks have also emerged focusing on document grounding for dialogue response generation (Zhou et al., 2018; Dinan et al., 2018). Zhao et al. (2020a) explore this task in low-resource setting and use pre-training along with a disentangled decoder. The disentangled decoder consists of a context processor, knowledge processor and a language model. A dialogue manager is used to combine the vocabulary distributions provided by these three components. Zhao et al. (2020b) propose a knowledge selection module integrated with pre-trained language models for this task. Cao et al. (2020) use pre-trained language model GPT-2 (Radford et al.) and explore various attention fusion techniques for persona-based dialogue generation (Zhang et al., 2018; Dinan et al., 2020b). Our DoHA technique also introduces an additional attention multi-head but does not use any additional weights to fuse attention heads. Similarly, Junczys-Dowmunt and Grundkiewicz (2018) use an additional attention multi-head in transformer architecture for automatic post-editing task. We demonstrate how attention can be enhanced in pre-trained models. Although Bruyn et al. (2020) introduce the usage of BART for knowledge grounded dialogues, it is primarily from the perspective of improving knowledge retrieval. We provide benchmark BART numbers (Table 4.8) for the generation task. Prabhumoye et al. (2020a) provide a schema containing five modules which can be changed to control the generation process. While Zhao et al. (2020a) modify the external input and the output module, we focus on the external input and the generator module of the pre-trained language model. These techniques can be applied to update summarization (Dang and Owczarzak, 2008), which involves generating a summary of the novel information in a new article assuming the user has read prior documents.

## 4.6 Conclusion

This chapter highlights the importance of the task of document grounded generation: generation guided by an existing curated text to set context and tone, and grounded in a new source

providing useful information. We define two concrete tasks to study and explore document grounded generation: *Wikipedia Update Generation* and *Document grounded dialogue response generation*.

The setting of *Wikipedia Update Generation* is particularly promising given the opportunity for human interaction: in contrast to approaches that do not rely on human-generated context, the task establishes a collaboration between user and computer. Each newly suggested sentence can be rejected, accepted, or edited before inclusion, and the edits can provide more training data. We believe there are many natural extensions to this work. One could apply models iteratively to incorporate changes for a set of documents and generate longer text.

We propose two novel improvements for document grounded generation and provide a strong baseline. The proposed models outperform the previous techniques and the new stronger baseline on automated metrics and human evaluation for the three datasets discussed. We present a comprehensive manual inspection which reveals certain data artifacts and provides us with insight on how to model these tasks in future. Particularly, future work can focus on designing better evaluation metrics which don't penalize linguistic variations in generation. Better models can also be constructed to focus on cases of partial hallucination or incorrect responses.



## Chapter 5

# Sentence Ordering

In the Mesopotamian era writing began as a consequence of political expansion, which needed reliable means for transmitting information, maintaining financial accounts, keeping historical records, and similar activities. In the current millennia, writing serves multiple functions which include - improvised additional capacity for the limitations of human memory (e.g. recipes, reminders, logbooks, the proper sequence for a complicated task or important ritual), dissemination of ideas (as in an essay, manifesto etc), imaginative narratives and other forms of storytelling, personal or business correspondence, and lifewriting (e.g., a diary or journal). Note that all of these functions of writing ranging from the ancient cultures to the current day, need the ideas or narratives in the written text to be organized in a logically coherent structure. The arrangement of the words and sentences in a document come together to convey the purpose of the text. Our goal is to model this arrangement of text.

Language in the real world has structure. Written texts have constrained structures that vary by genre. Newspaper articles, for instance, typically have an “inverted pyramid” structure: a recent event, followed by the most relevant details of that event, followed by secondary background information at the end of the article. Wikipedia articles, by contrast, are often chronological, beginning with the earliest major historical event on a topic and proceeding sequentially. Various frameworks have been designed to understand document structures. Document structures have been modeled as trees based on the relations between the sentences in Rhetorical Structure Theory (RST; Mann and Thompson (1988)), as graphs (Wolf and Gibson, 2006), and as entity grid model based on transitions (Barzilay and Lapata, 2008). Each of these frameworks capture different aspects of structure - structure between two consecutive sentences is captured by local coherence, the relation of a sentence with other sentences of the document captures global structure as in the case of RST, structure can be captured between the ordering of the paragraphs in the document. Structure could also mean the ordering of events in a document, the transition of topics in a document or dialogue, the transition of scenes or a plot in a story, the chronology of events in a Wikipedia article etc.

Document structures are useful in modeling and interpreting various Natural Language Processing tasks. They are important for summarization (Barzilay and McKeown, 2005), automated essay scoring (Burstein et al., 2010; Miltsakaki and Kukich, 2004), question-answering (Verberne et al., 2007), text planning (Hovy, 1988; Marcu, 1997) and document classification (Liu and Lapata, 2018). We primarily care about the sentence ordering sub-task which is important to understand document structures.

Sentence ordering is the task of arranging sentences into an order which maximizes the coherence of the text (Barzilay and Lapata, 2008). This subtask provides us insights into modeling the ordering of events in a document. Recent work has modeled this task as a sequence generation task using hierarchical neural models. We have framed the task as a constraint learning problem. We train a model which learns to predict the correct constraint given a pair of sentences. The constraint learnt by our model is the relative ordering between the two sentences. Given a set of constraints between the sentences of a document, we find the right order of the sentences by using sorting techniques. Simple sorting techniques can outperform the previous approaches by a large margin given that it has good sentence representations. The bottleneck for most of the hierarchical models is memory required by the representations of all the sentences and the representation of the paragraph. The new framing also obviates these memory issues.

**Overview** : We present the new framing of the sentence ordering task in §5.1. The experiments and results are showcased in §5.2 and §5.3. We present extensive analysis of the results in §5.3.1 and a literature survey in §5.4. This work is done in collaboration with Alan W Black and Ruslan Salakhutdinov (Prabhumoye et al., 2020c).

## 5.1 Methodology

For this task we have a set of  $N$  documents  $\mathcal{D} = \{d_1, \dots, d_N\}$ . Let the number of sentences in each document  $d_i$  be denoted by  $v_i$ , where  $\forall i, v_i \geq 1$ . The task can be formulated as - If you have a set  $\{s_{o_1}, \dots, s_{o_{v_i}}\}$  of  $v_i$  sentences in a random order where the random order is  $\mathbf{o} = [o_1, \dots, o_{v_i}]$ , then the task is to find the right order of the sentences  $\mathbf{o}^* = [o_1^*, \dots, o_{v_i}^*]$ . Prior work (Logeswaran et al., 2018b; Cui et al., 2018) learns to predict the sequence of the correct order  $\mathbf{o}^*$ . In this formulation of the task, we have  $\mathcal{C}_i$  set of constraints for document  $d_i$ . These constraints  $\mathcal{C}_i$  represent the relative ordering between every pair of sentences in  $d_i$ . Hence, we have  $|\mathcal{C}_i| = \binom{v_i}{2}$ . For example, if a document has four sentences in the correct order  $s_1 < s_2 < s_3 < s_4$ , then we have six set of constraints  $\{s_1 < s_2, s_1 < s_3, s_1 < s_4, s_2 < s_3, s_2 < s_4, s_3 < s_4\}$ . Constraints  $\mathcal{C}_i$  are learnt using a classifier neural network described in (§5.1.2). We finally find the right order  $\mathbf{o}^*$  using topological sort on the relative ordering between all the  $\mathcal{C}_i$  pairs of sentences.

### 5.1.1 Topological Sort

Topological sort (Tarjan, 1976) is a standard algorithm for linear ordering of the vertices of a directed graph. The sort produces an ordering  $\hat{o}$  of the vertices such that for every directed edge  $u \rightarrow v$  from vertex  $u$  to vertex  $v$ ,  $u$  comes before  $v$  in the ordering  $\hat{o}$ . We use the depth-first search based algorithm which loops through each node of the graph, in an arbitrary order. The algorithm visits each node  $n$  and prepends it to the output ordering  $\hat{o}$  only after recursively calling the topological sort on all descendants of  $n$  in the graph. The algorithm terminates when it hits a node that has been visited or has no outgoing edges (i.e. a leaf node). Hence, we are guaranteed that all nodes which depend on  $n$  are already in the output ordering  $\hat{o}$  when the algorithm adds node  $n$  to  $\hat{o}$ .

We use topological sort to find the correct ordering  $\mathbf{o}^*$  of the sentences in a document. The sentences can represent the nodes of a directed graph and the directed edges are represented by the ordering between the two sentences. The direction of the edges are the constraints predicted by the classifier. For example, if the classifier predicts the constraint that sentence  $s_1$  precedes  $s_2$ , then the edge  $s_1 \rightarrow s_2$  would be from node of  $s_1$  to  $s_2$ .

This algorithm has time complexity of  $O(v_i + |\mathcal{C}_i|)$  for a document  $d_i$ . In our current formulation, all the constraints are predicted before applying the sort. Hence, we have to consider all the  $|\mathcal{C}_i| = \binom{v_i}{2}$  edges in the graph. The time complexity of our current formulation is  $O(v_i^2)$ . But the same technique could be adopted using a Merge Sort (Knuth, 1998) algorithm in which case the time complexity would be  $O(v_i \log v_i)$ . In this case, the sort algorithm is applied first and the constraint is predicted only for the two sentences for which the relative ordering is required during the sort time.

### 5.1.2 Constraint Learning

We build a classifier to predict a constraint between two sentences  $s_1$  and  $s_2$  (say). The constraint learnt by the classifier is the relative ordering between the two sentences. Specifically, the classifier is trained to predict whether  $s_2$  follows  $s_1$  or not i.e the classifier predicts the constraint  $s_1 < s_2$ .

**BERT based Representation (B-TSort):** We use the Bidirectional Encoder Representations from Transformers (BERT) pre-trained uncased language model (Devlin et al., 2019) and fine-tune it on each dataset using a fully connected perceptron layer. Specifically, we leverage the Next Sentence Prediction objective of BERT and get a single representation for both sentences  $s_1$  and  $s_2$ . The input to the BERT model is the sequence of tokens of sentence  $s_1$ , followed by the separator token '[SEP]', followed by the sequence of tokens for sentence  $s_2$ . We use the pooled representation for all the time steps.

Dataset	Length Statistics			Data split			Vocabulary
	min	mean	max	train	valid	test	
NIPS	2	6.0	15	2248	409	402	16721
AAN	1	5.0	20	8569	962	2626	34485
NSF	2	8.9	40	96070	10185	21580	334090
SIND	5	5.0	5	40155	4990	5055	30861

TABLE 5.1: Dataset Statistics

**LSTM based Representation (L-TSort):** In this model we get two separate representations  $\mathbf{h}_1$  and  $\mathbf{h}_2$  for  $s_1$  and  $s_2$  from a bi-directional LSTM encoder, respectively. We pass the concatenation of  $\mathbf{h}_1$  and  $\mathbf{h}_2$  as input to a two layers of perceptron for constraint prediction. This model is trained to gain insight on the contribution of pre-trained sentence representations for the constraint prediction formulation of the task.

## 5.2 Experiments

This section describes the datasets, the evaluation metric and the results of our experiments.

### 5.2.1 Datasets

The dataset statistics for all the datasets are shown in Table 5.1.

**NSF, NIPS, AAN abstracts:** These three datasets contain abstracts from NIPS papers, ACL papers, and the NSF Research Award Abstracts dataset respectively and are introduced in (Logeswaran et al., 2018b). The paper also provides details about the statistics and processing steps for curating these three datasets.

**SIND caption:** We also consider the SIND (Sequential Image Narrative Dataset) caption dataset (Huang et al., 2016) used in the sentence ordering task by (Gong et al., 2016). All the stories in this dataset contain five sentences each and we only consider textual stories for this task.

### 5.2.2 Baselines

**Attention Order Network (AON):** This is the current state-of-the-art model (Cui et al., 2018) which formulates the sentence ordering task as a order prediction task. It uses a LSTM based encoder to learn the representation of a sentence. It then uses a transformer network based paragraph encoder to learn a representation of the entire document. It then decodes the sequence of the order by using a LSTM based decoder.

**BERT Attention Order Network (B-AON).** To have a fair comparison between our model and the AON model, we replace the LSTM based sentence representation with the pre-trained uncased BERT model. This model plays a pivotal role of giving us an insight into how much improvement in performance we get only due to BERT.

### 5.2.3 Evaluation Metric

**Perfect Match (PMR):** It is the strictest metric and calculates the percentage of samples for which the entire sequence was correctly predicted (Chen et al., 2016).  $PMR = \frac{1}{N} \sum_{i=1}^N 1\{\hat{\mathbf{o}}^i = \mathbf{o}^{*i}\}$ , where  $N$  is the number of samples in the dataset. This is the strictest metric and gives us an absolute accuracy of the whole order being predicted correctly.

**Sentence Accuracy (Acc):** It is a stringent metric and measures the percentage of sentences for which their absolute position was correctly predicted (Logeswaran et al., 2018b).  $Acc = \frac{1}{N} \sum_{i=1}^N \frac{1}{v_i} \sum_{j=1}^{v_i} 1\{\hat{\mathbf{o}}_j^i = \mathbf{o}_j^{*i}\}$ , where  $v_i$  is the number of sentences in the  $i^{th}$  document. This is a stringent metric and tells us the percent of sentences within a document that we can predict the right order for.

**Kendall Tau (Tau):** This metric quantifies the distance between the predicted order and the correct order in terms of the number of inversions (Lapata, 2006). It calculates the number of inversions required by the predicted order to reach the correct order.  $\tau = 1 - 2I/\binom{v_i}{2}$ , where  $I$  is the number of pairs in the predicted order with incorrect relative order and  $\tau \in [-1, 1]$ .

**Rouge-S (R-S):** It calculates the percentage of skip-bigrams for which the relative order is predicted correctly (Chen et al., 2016). Skip-bigrams are the total number of pairs  $\binom{v_i}{2}$  in a document. Note that it does not penalize any arbitrary gaps between two sentences as long as their relative order is correct.  $Rouge-S = \frac{1}{\binom{v_i}{2}} \text{Skip}(\hat{\mathbf{o}}) \cap \text{Skip}(\mathbf{o}^*)$ , where the  $\text{Skip}(\cdot)$  function returns the set of skip-bigrams of the given order.

**Longest Common Subsequence (LCS):** It calculates the ratio of longest correct sub-sequence (Gong et al., 2016) (consecutiveness is not necessary, and higher is better).

**Human Evaluation** We introduce a human evaluation experiment to assess the orders predicted by the models. We set up a manual pairwise comparison following (Bennett, 2005) and present the human judges with two orders of the same piece of text. The judges are asked ‘Pick the option which is in the right order according to you.’ They can also pick a third option ‘No Preference’ which corresponds to both the options being equally good or bad. In total we had

Model	PMR	Acc	Tau	Rouge-S	LCS	PMR	Acc	Tau	Rouge-S	LCS
NIPS abstracts					SIND captions					
AON	16.25	50.50	0.67	80.97	74.38	13.04	45.35	0.48	73.76	72.15
B-AON	19.90	55.23	0.73	83.65	76.29	14.30	47.73	0.52	75.77	73.48
L-TSort	12.19	43.08	0.64	80.08	71.11	10.15	42.83	0.47	73.59	71.19
B-TSort	<b>32.59</b>	<b>61.48</b>	<b>0.81</b>	<b>87.97</b>	<b>83.45</b>	<b>20.32</b>	<b>52.23</b>	<b>0.60</b>	<b>78.44</b>	<b>77.21</b>

TABLE 5.2: Results on five automatic evaluation metrics for NIPS and SIND datasets.

Model	PMR	Acc	Tau	Rouge-S	LCS	PMR	Acc	Tau	Rouge-S	LCS
NSF abstracts					AAN abstracts					
AON	13.18	38.28	0.53	69.24	61.37	36.62	56.22	0.70	81.52	79.06
B-TSort	10.44	35.21	0.66	69.61	68.50	50.76	69.22	0.83	87.76	85.92

TABLE 5.3: Results on five evaluation metrics for NSF and AAN datasets.

100 stories from the SIND dataset<sup>1</sup> annotated by 10 judges. We setup three pairwise studies to compare the B-TSort vs AON order, B-TSort vs Gold order and AON vs Gold order (Gold order is the actual order of the text).

### 5.3 Results

Table 5.2 shows the results of the automated metrics for the NIPS and SIND datasets<sup>2</sup>. It shows that AON<sup>3</sup> model gains on all metrics when the sentence embeddings are switched to BERT. The L-TSort model which does not utilize BERT embeddings comes close to AON performance on Rouge-S and Tau metrics. This demonstrates that the simple L-TSort method is as accurate as AON in predicting relative positions but not the absolute positions (PMS and Acc metric). Table 5.2 shows that our method B-TSort does not perform better only due to BERT embeddings but also due to the experiment design. Note that BERT has been trained with the Next Sentence Prediction objective and not the sentence ordering objective like ALBERT (Lan et al., 2019). We believe that framing this task as a constraint solving task will further benefit from pre-trained language model like ALBERT. Table 5.3 shows results for the NSF and AAN datasets and the B-TSort model performs better than the AON model on all metrics.

Table 5.4 shows results for the three human evaluation studies on the SIND dataset. It shows that human judges prefer B-TSort orders 10% more number of times than the AON orders. The reference order may not be the only correct ordering of the story. The variability in the orders

<sup>1</sup>We choose SIND because all the stories contain 5 sentences and hence it is easy to read for the judges. The orders of the stories are easier to judge as compared to the orders of scientific abstracts like NSF, NIPS and AAN as they require the judges to have an informed background.

<sup>2</sup>We fine-tune BERT which is memory intensive. Hence, we show the results of B-AON only on these two datasets as they need 2 transformer layers for paragraph encoder (Cui et al., 2018)

<sup>3</sup>We use the code provided by the authors to train the AON and B-AON model. The numbers reported in Table 5.2 and 5.3 are our runs of the model. Hence, they differ from the numbers reported in the paper (Cui et al., 2018).

B-TSort vs B-AON			B-TSort vs Gold			B-AON vs Gold		
B-TSort	No Pref	B-AON	B-TSort	No Pref	Gold	B-AON	No Pref	Gold
<b>41.00%</b>	28.00%	31.00%	26.00%	20.00%	<b>54.00%</b>	24.00%	22.00%	<b>54.00%</b>

TABLE 5.4: Human Evaluation Results on B-TSort vs AON (top), B-TSort vs Gold (middle) and AON vs Gold (bottom).

Model	Win=1	Win=2	Win=3	% Miss	Win=1	Win=2	Win=3	% Miss
<b>NIPS</b>					<b>SIND</b>			
B-AON	81.81	92.44	96.50	3.48	78.39	92.79	98.43	0.00
B-TSort	87.59	95.59	98.11	0.00	82.67	95.01	99.09	0.00
<b>NSF</b>					<b>AAN</b>			
AON	50.58	63.87	72.96	5.85	82.65	92.25	96.73	0.84
B-TSort	61.41	75.52	83.87	0.00	90.56	96.78	98.71	0.00

TABLE 5.5: Displacement Analysis for all the datasets.

produced by B-TSort and AON is not very high and hence in comparison with Gold orders, we don't see much difference in human preferences.

The low scores of AON could be due to the fact that it has to decode the entire sequence of the order. The search space for decoding is very high (in the order of  $v_i!$ ). Since our framework, breaks the problem to a pairwise constraint problem, the search space for our model is in the order of  $v_i^2$ .

### 5.3.1 Discussion

We perform a few additional experiments to determine the displacement of sentences in the predicted orders by B-TSort model due to lack of direct global structure, scalability of the model for documents containing more than ten sentences, and an understanding of quality of the human judgements.

To understand the displacement of sentences in the predicted orders, we calculate the percentage of sentences whose predicted location is within 1, 2 or 3 positions (in either direction) from its original location. We observed that B-TSort consistently performs better on all datasets for all three window sizes as shown in Table 5.5. Observe that as window size reduces, the difference between B-TSort and B-AON percentages increases. This implies that displacement of sentences is higher in B-AON despite taking the whole document into account.

We additionally perform a comparison of models on documents containing more than 10 sentences and the results are shown in Table 5.6. B-TSort consistently performs better on all the metrics. SIND dataset is omitted in these experiments as the maximum number of sentences in the story is five for all the stories in the dataset. Note that the AON model generates the order and hence need not generate positions for all the sentences in the input. We calculate the

<b>Model</b>	<b>PMR</b>	<b>Acc</b>	<b>Tau</b>	<b>Rouge-S</b>	<b>LCS</b>	<b>%Mismatch</b>
<b>NIPS abstracts</b>						
B-AON	0.0	29.18	0.51	74.64	63.81	33.33
B-TSort	0.0	39.43	0.74	83.26	71.68	0.00
<b>NSF abstracts</b>						
AON	2.12	21.42	0.41	67.45	55.47	11.60
B-TSort	0.67	28.57	0.64	68.46	64.86	0.00
<b>AAN abstracts</b>						
AON	0.0	22.70	0.40	68.90	56.19	5.17
B-TSort	0.0	36.86	0.69	78.52	72.01	0.00

TABLE 5.6: Analysis on NIPS, NSF and AAN datasets on documents longer than 10 sentences.

percentage of mismatches between the length of the input document and the generated order. For NSF dataset, the overall mismatch is 3.48%, while the mismatch for documents with more than 10 sentences is 11.60% as shown in Table 5.6. This problem does not arise in our design of the task as it does not have to stochastically generate orders.

To better understand the choices of human judges, we observe the average length of stories calculated in number of tokens. We discover that the average length of the stories is 86 for B-TSort which is much higher as compared to B-AON with average length of 65. The average length of stories is 47 when 'No Preference' option is chosen for B-TSort vs B-AON. This means that B-TSort is better according to human judges for longer stories. Similarly for B-TSort vs Gold experiment, the human judges were confused with longer stories, reiterating that B-TSort performs well with long stories.<sup>4</sup>

## 5.4 Related Work

Sentence ordering is the task of arranging sentences into an order which maximizes the coherence of the text (Barzilay and Lapata, 2008). This is important in applications where we have to determine the sequence of pre-selected set of information to be presented. This task has been well-studied in the community due to its significance in down stream applications such as ordering of: concepts in concept-to-text generation (Konstas and Lapata, 2012), information from each document in multi-document summarization (Barzilay and Elhadad, 2002; Nallapati et al., 2017), events in storytelling (Fan et al., 2019; Hu et al., 2020a), cooking steps in recipe generation (Chandu et al., 2019a), and positioning of new information in existing summaries for update summarization (Prabhumoye et al., 2019b). Student essays are evaluated based on how coherent and well structured they are. Hence, automated essay scoring (Burstein et al., 2010; Miltsakaki and Kukich, 2004) can use this task to improve the efficiency of their systems.

<sup>4</sup>Appendix C details the hyper-parameters used for both the models and presents examples of the orders predicted for SIND and NIPS datasets by the two models.



Early work on coherence modeling and sentence ordering task uses probabilistic transition model based on vectors of linguistic features (Lapata, 2003), content model which represents topics as states in an HMM (Barzilay and Lee, 2004), and entity based approach (Barzilay and Lapata, 2008). Recent work uses neural approaches to model coherence and to solve sentence ordering task. Li and Hovy (2014) introduced a neural model based on distributional sentence representations using recurrent or recursive neural networks and avoided the need of feature engineering for this task. In (Li and Jurafsky, 2017), they extend it to domain independent neural models for coherence and they introduce new latent variable Markovian generative models to capture sentence dependencies. These models used windows of sentences as context to predict sentence pair orderings. Gong et al. (2016) proposed end-to-end neural architecture for sentence ordering task which uses pointer networks to utilize the contextual information in the entire piece of text.

Recently hierarchical architectures have been proposed for this task. In (Logeswaran et al., 2018b), the model uses two levels of LSTMs to first get the encoding of the sentence and then get the encoding of the entire paragraph. Cui et al. (2018) use a transformer network for the paragraph encoder to allow for reliable paragraph encoding. Prior work (Logeswaran et al., 2018b; Cui et al., 2018) has treated this task as a sequence prediction task where the order of the sentences is predicted as a sequence. The decoder is initialized by the document representation and it outputs the index of sentences in sequential order. Only in (Chen et al., 2016), this task is framed as a ranking problem. In this work, a pairwise score is calculated between two sentences and then the final score for an order is obtained by summing over all the scores between pairs of sentences. The order which has the maximum score is given as output. Instead of considering all possible permutations of a given order, it uses beam-search strategy to find a sub-optimal order.

Most of the recent work (Gong et al., 2016; Logeswaran et al., 2018b; Cui et al., 2018) tries to leverage the contextual information but has the limitation of predicting the entire sequence of the order. This has the limitation that the prediction at the current time step is dependent on the prediction of the previous time step. Another limitation of the prior work is the availability of good sentence representations that can help in determining the relative order between two sentences.

## 5.5 Conclusion

We have shown a new way to design the task of sentence ordering. We provide a simple yet efficient method to solve the task which outperforms the state of the art technique on all metrics. We acknowledge that our current model has the limitation of not including the entire context of the paragraph while making the decision of the relative order of the pairs. The future work can include the paragraph representation in the constraint prediction model. This will help the methodology to have the benefit of making informed decisions while also solving constraints.

## Chapter 6

# Ethical Considerations

The 21st century is witnessing a major shift in the way people interact with technology, and natural language processing (NLP) is playing a central role. A plethora of NLP applications such as question-answering systems (Bouziane et al., 2015; Gillard et al., 2006; Yang et al., 2018a) used in diverse fields like health care (Sarrouiti and Ouatik El Alaoui, 2017; Zweigenbaum, 2009), education (Godea and Nielsen, 2018; Raamadhurai et al., 2019), privacy (Ravichander et al., 2019; Shvartzshanider et al., 2018); machine translation systems (Cherry et al., 2019; Barrault et al., 2019; Nakazawa et al., 2019; Liu, 2018), conversational agents (Pietquin et al., 2020; Serban et al., 2018; Liu et al., 2016), recommendation systems (Alharthi and Inkpen, 2019; Greenquist et al., 2019) etc are deployed and used by millions of users. NLP systems have become pervasive in current human lifestyle by performing mundane tasks like setting reminders and alarms to complex tasks like replying to emails, booking tickets and recommending movies/restaurants. This widespread use calls for an analysis of these systems from an ethical standpoint.

Despite all the advances in efficiency and operations of NLP systems, little literature exists which broadly addresses the ethical challenges of these technologies. Ethical theories have been studied for millennia and should be leveraged in a principled way to address the questions we are facing in NLP today. Instead, the topic of “ethics” within NLP has come to refer primarily to addressing bias in NLP systems; Blodgett et al. (2020) provides a critical survey of how “bias” is studied in NLP literature. The survey finds that research on NLP systems conceptualize “bias” differently and that the techniques are not well tied with the relevant literature outside of NLP. This creates a gap between NLP research and the study of ethics in philosophy which leaves a rich body of knowledge untapped.

This chapter bridges this gap by illustrating how a philosophical theory of ethics can be applied to NLP research. Ethics (or ethical theory), most broadly, is a theoretical and applied branch of philosophy which studies what is good and right, especially as it pertains to how humans *ought* to behave in the most general sense (Fieser, 1995). As NLP research qualifies as a human activity, it is within the purview of ethics. In particular, we are using a *prescriptive*, rather

than *descriptive*, theory of ethics; prescriptive theories define and recommend ethical behavior whereas descriptive theories merely report how people generally conceive of ethical behavior.

We select two ethical principles from the deontological tradition of ethics and focus on how these principles are relevant to research in NLP. Namely we look at the *generalization principle* and *respect for autonomy* through informed consent (Johnson and Cureton, 2019; Kleinig, 2009). We select deontology because it is reasonable, provides clear ethical rules and comports with the legal idea of the *rule of law* in the sense that these ethical rules bind all persons equally, rather than addressing situations in a case-by-case basis.

We find that there are two main ways in which ethical guidelines can be applied in NLP (or to any other area of technology):

1. An ethical guideline can aid in deciding *what* topics within a field merit attention; that is, it answers the question “which tasks have important ethical implications?”.
2. An ethical guideline can aid in determining *how* to address a problem; that is, it answers the question “what factors and methods are preferable in ethically solving this problem?”.

We primarily address (1) and briefly touch on (2) by presenting four case studies relevant to NLP. In each case study we use an ethical principle to identify an area of research that could potentially conflict with it, and suggest NLP directions to mitigate it. Note that some issues identified in this chapter may be well known and yet it is important to analyze them through the lens of a systematic framework. We believe that research should not be subjective and should be based on formal methods which are consistent and the judgements are reproducible. Additionally, this chapter identifies practical NLP methods that are supported by the chosen ethical principles in resolving the issues discussed.

Although we have selected two principles from a deontological perspective, we are not intimating that these principles can address all ethical issues nor that deontological ethics is the only ethical framework in which our rules and case studies could function (§6.5). Instead, we present the following as a starting point for NLP researchers less familiar but interested in applicable ethical theory.

**Relevance to Controllable Text Generation:** Since NLP systems interact heavily with humans, sometimes making decisions on behalf of the humans, we need to be aware of the ethical issues of these systems. Most of the NLP systems use controllable text generation to perform some of their functions like response generation in conversational agents, machine translation, and generating answers in Question-Answering systems. As we have discussed in earlier chapters (§3.3 and §4.3.3), there is a need of better evaluation of the current NLP systems. In this chapter we show that *ethics* can be considered a separate dimension along which NLP systems can be evaluated. In the following sections, we see that controllable text generation techniques are important in ensuring that NLP systems cater to all demographics

and groups of people. Furthermore, we observe that controllable text generation techniques can be used to make systems ethical (§6.3).

The chapter provides an overview of two deontological principles along with a discussion on their limitations with a special focus on NLP. It illustrates four specific case studies of NLP systems which have ethical implications under these principles and providing a direction to alleviate these issues.

**Overview:** We first provide a comprehensive literature survey of both work done in the field of ethics and ethics in NLP (§6.1). We then describe in detail the three principles from deontological ethics with examples (§6.2). In §6.3, we present four case studies and demonstrate how these ethical principles can be applied to real world NLP applications. We also further identify practical NLP methods that can be used to mitigate the ethical challenges in "The way forward" sections of each of the case studies. Furthermore, we also present how ethical decision making can be aided by NLP tools in §6.4. Finally, we present a discussion on the limitations of the choice of ethical principles and future directions (§6.5). Most of the work presented in this chapter was done in collaboration with Brendon Boldt, Alan W Black and Ruslan Salakhutdinov (§6.1, §6.2, §6.3, §6.4, and §6.5) and a part of it was done in collaboration with Elijah Mayfield (§6.2.3 and §6.5).

## 6.1 Related Work

### 6.1.1 Ethics

While there are a number of categories of prescriptive ethical theories, including deontology (Kant, 1785), consequentialism (e.g., utilitarianism) (Bentham, 1843), and virtue ethics (Aristotle, 350 B.C.E.), we are only addressing deontology. We do not take a stance in this paper as to whether or not there exists an objectively correct ethical theory, but we offer a brief sketch of deontological ethics and our reasons for using it. Deontology or deontological ethics refers to a family of ethical theories which hold that whether an act is ethically good or bad is determined by its adherence to ethical rules (Alexander and Moore, 2016). These rules can be agent-focused duties (e.g., duty to care for one's children) or patient-focused rights (e.g., right to life). Such rules can also be formulated in modal logic, allowing for more precise reasoning over sets of rules (Hooker and Kim, 2018).

Deontology stands in contrast to another popular framework of ethics: consequentialism. Consequentialism holds the ultimate consequences of an action to be the deciding factor regardless of the nature of the actions taken to get there. We can illustrate the difference between them by observing how each of them might condemn something like racially biased hiring in academia.<sup>1</sup> A deontologist might say that this practice is wrong because it violates the human

---

<sup>1</sup>Note that we are presenting generic examples of deontological and consequentialist frameworks and that a variety of nuanced theories in each category exist.

right to equal treatment regardless of race. A consequentialist on the other hand, would argue that this is wrong because its *effect* is stymieing academic creativity by reducing intellectual diversity.

We ultimately select the deontological framework in this work for the following reasons:

1. We find deontology to be convincing in its own right, namely, its ability to delineate robust duties and rights which protect the value of each and every person.
2. The universally applicable rules of deontology provide a good basis for providing recommendations to researchers. Since rights and duties (at their core) are not situation dependent, they are tractable to address in NLP applications.<sup>2</sup>
3. The focus on rights and duties which apply to everyone equally fits well with the widespread legal concept of the *rule of law* which states that every person is subject to the same laws.

### 6.1.2 Ethics in NLP

We appeal to the fact that problems should be analyzed with a systematic framework, and ethical theories provide precisely these frameworks. Research should not be based on preconceived notions of ethics which can be overly subjective and inconsistent. To more rigorously determine what is right and wrong, we rely on ethical theories. [Card and Smith \(2020\)](#) present an analysis of ethics in machine learning under a consequentialist framework. This paper is a kindred spirit in that both seek to make a philosophical theory of ethics concrete within machine learning and NLP, yet the methods of the paper are somewhat orthogonal. [Card and Smith \(2020\)](#) provide a comprehensive overview of how the particular nature of consequentialist ethics is relevant to machine learning whereas we intend to provide tangible examples of how deontological ethical principles can identify ethically important areas of research. [Saltz et al. \(2019\)](#); [Bender et al. \(2020\)](#) advocate for explicitly teaching ethical theory as a part of machine learning and NLP courses; the case studies in this paper would be a logical extension of the material presented in such a course.

NLP research has primarily focused on two directions: (1) exploring and understanding the impact of NLP on society, and (2) providing algorithmic solutions to ethical challenges.

[Hovy and Spruit \(2016\)](#) started the conversation about the potential social harms of NLP technology. It discussed the concepts of *exclusion*, *overgeneralization*, *bias confirmation*, *topic under- and overexposure*, and *dual use* from the perspective of NLP research. A lot of work followed this discussion and made contributions towards ethical frameworks and design practices ([Leidner and Plachouras, 2017](#); [Parra Escartín et al., 2017](#); [Prabhumoye et al., 2019a](#); [Schnoebelen, 2017](#); [Schmaltz, 2018](#)), data handling practices ([Lewis et al., 2017](#); [Mieskes, 2017](#)) and specific domains like education ([Mayfield et al., 2019](#); [Loukina et al., 2019](#)), health care ([Šuster et al.,](#)

<sup>2</sup>In contrast to (action-based) utilitarianism which mandates evaluating the full consequences of each action.

2017; Benton et al., 2017) and conversational agents (Cercas Curry and Rieser, 2018; Henderson et al., 2018). Our paper does not focus on a particular domain but calls for attention towards various NLP systems and what ethical issues may arise in them.

Most of the work providing algorithmic solutions has been focused on bias in NLP systems. Shah et al. (2020); Tatman (2017); Larson (2017) aim to study the social impact of bias in NLP systems and propose frameworks to understand it better. A large body of work (Bolukbasi et al., 2016; Sun et al., 2019; Zhao et al., 2019, 2017; Sap et al., 2019; Hanna et al., 2020; Davidson et al., 2019) directs its efforts to mitigate bias in data, representations, and algorithms. Blodgett et al. (2020) provide an extensive survey of this work and point out the weaknesses in the research design. It makes recommendations of grounding work analyzing “bias” in NLP systems in the relevant literature outside of NLP, understanding why system behaviors can be harmful and to whom, and engaging in a conversation with the communities that are affected by the NLP systems. Although issues with bias are certainly within the scope of the principles we present, we do not specifically write on bias because it has already received a large amount of attention.

## 6.2 Deontological Ethics

There are a variety of specific deontological theories which range from having one central principle (Kant, 1785) to having a handful of concrete principles (Ross, 1930). Rather than comprehensively addressing one theory, we select two rules, one abstract and one concrete, which can fit within a variety of deontological theories. The *generalization principle* is an abstract, broad-reaching rule which comes from traditional Kantian ethics. The *respect for autonomy* is concrete and commonly seen in political and bioethics.

### 6.2.1 Generalization Principle

The generalization principle has its roots in Immanuel Kant’s theory of deontological ethics (Kant, 1785).<sup>3</sup> The generalization principle states the following (Johnson and Cureton, 2019).

An action  $\mathcal{A}$  taken for reasons  $\mathcal{R}$  is ethical if and only if a world where all people perform  $\mathcal{A}$  for reasons  $\mathcal{R}$  is conceivable.

It is clearer when phrased in the negative.

An action  $\mathcal{A}$  taken for reasons  $\mathcal{R}$  is *unethical* if and only if a world where all people perform  $\mathcal{A}$  for reasons  $\mathcal{R}$  logically contradicts  $\mathcal{R}$ .

---

<sup>3</sup>It is also referred to as the “universal law” formulation of Kant’s categorical imperative.

The main utility of the generalization principle is that it can identify unethical actions that may seem acceptable in isolated occurrences but lead to problems when habitually taken by everyone.

This approach is founded on the work of (Kant, 1785), which fundamentally prioritizes *intent* as the source of ethical action. To analyze this in machine learning, a trained agent  $\mathcal{G}$  is expected to take an action  $\mathcal{A}_i$  based on a given set of evidence  $\mathbf{E}_i$ , from a finite closed set of options  $\mathcal{A}$ . This simple notation can be extended to classification, regression, or reinforcement learning tasks. The generalization principle states that agent  $\mathcal{G}$  is ethical if and only if, when given two identical sets of evidence  $\mathbf{E}_1$  and  $\mathbf{E}_2$  with the *same* inputs, agent  $\mathcal{G}$  chooses to make same decision  $\mathbf{A}_1$  every time. Furthermore, the principle assumes that all *other* such trained agents will *also* make those same predictions.

For example, let us take making and breaking a legal contract (the action) whenever it is convenient (the reasons); implicit in the reasons for making a contract is that the other person believes we will follow through (Johnson and Cureton, 2019). If we universalize this and conceive of a world where everyone makes contracts which they have no intent of keeping, no one would believe in the sincerity of a contract. Hence, no one would make contracts in the first place since they are never adhered to. This is the sort of contradiction by which the generalization principle condemns an action and the rationale behind it.

Another example is plagiarism of research papers in conference submissions. Let us assume that a top tier conferences did not check for plagiarism because they trust in the honesty of the researchers. In this case, a researcher  $\mathbf{G}$  decides to take an action  $\mathcal{A}$  of plagiarising a paper due to the following set of reasons  $\mathcal{R}$ : (1)  $\mathbf{G}$  believes that they would not get caught because the conference does not use plagiarism detection software, (2) publishing this paper in the said conference would boost  $\mathbf{G}$ 's profile by adding 100 citations, and (3) this would increase  $\mathbf{G}$ 's chances of getting a job. Plagiarism in this case would be ungeneralizable and hence unethical. If all researchers who want to boost their profile were to submit plagiarised papers, then every researcher's profile would be boosted by 100 citations, and 100 citations would lose their value. Hence, this would not increase  $\mathbf{G}$ 's chances of getting a job, contradicting  $\mathcal{R}$ 3. Thus,  $\mathbf{G}$ 's reasons for plagiarism are inconsistent with the assumption that everyone with same reasons plagiarises.

Here, we presume that the input representation is sufficient to make a prediction, without including any extraneous information. The reasons for an act define the *scope* of the act, or the set of necessary and sufficient conditions under which that act is generalizably moral (Hooker, 2018). Evidence must be relevant to the decision making process, and more-so must exclude task-irrelevant evidence that might be a source of bias. By excluding such evidence, the agent is invariant to *who* is being evaluated, and instead focuses its decision solely on task-relevant evidence.

This goal cannot be met without transparent and sparsely weighted inputs that do not use more information than is necessary and task-relevant for making predictions. Practically, this definition would privilege research on interpretable, generalizable, and understandable machine learning classifiers. The burden of proof of ethics in such a framework would lie on transparency and expressiveness of inputs, and well-defined, expected behavior from architectures for processing those features. Some work on this - like that from (Hooker and Kim, 2018) - has already begun. If deontological ethics were prioritized, we would expect to see rapid improvement in parity of  $F_1$  scores across subgroups present in our training data - an outcome targeted by practitioners like (Chouldechova, 2017) and (Corbett-Davies et al., 2017).

### 6.2.2 Respect for Autonomy

Respect for autonomy generally addresses the right of a person to make decisions which directly pertain to themselves. One of the primary manifestation of this is the concept of *informed consent*, whereby a person **A** proposes to act in some way  $\mathbb{X}$  on person **B** which would normally infringe on **B**'s right to self-govern. Specifically, we use the formulation of informed consent given by Pugh (2020) based on Kleinig (2009):

1. **B** must be sufficiently informed with regards to the relevant facts concerning  $\mathbb{X}$  to understand what  $\mathbb{X}$  is (and what consequences are likely to occur as a result of  $\mathbb{X}$ ).
2. On the basis of this information, **B** *herself* makes the decision to allow **A** to do  $\mathbb{X}$ .

Informed consent is of an important idea in bioethics where it typically applies to a patient's right to refuse treatment (or certain kinds of treatment) by medical personnel. In routine medical treatments this informed consent might be implicit, since one would not go to the doctor in the first place if they did not want to be treated at all, but in risky or experimental medical procedures, explaining the risks and benefits and obtaining explicit consent would be mandatory. In this case, the patient's autonomy specifically refers to opting out of medical procedures, and informed consent is a concrete method by which to respect this autonomy.

A non-medical example of respect for autonomy and informed consent would be hiring an interpreter **A** for a language that the user **B** does not speak. Under normal circumstances, **B**'s autonomy dictates that she and only she can speak for herself. But if she is trying to communicate in a language she does not speak, she might consent to **A** serving as an *ad hoc* representative for what she would like to say. In a high-stakes situation, there might be a formal contract of how **A** is to act, but in informal circumstances, she would *implicitly* trust that **A** translates what she says faithfully ( $\mathbb{X}$ ). In these informal settings, **A** should provide necessary information to **B** before deviating from the expected behaviour  $\mathbb{X}$  (e.g., if the meaning of a sentence is impossible to translate). Implicit consent is a double-edged sword: it is necessary to carry on normal social situations, but it can undermine the respect for autonomy in scenarios when (1) the person in question is not explicitly informed and (2) reasonable expectations do not match reality.



### 6.2.3 Utilitarian Principle

*An action is ethical only if it is not irrational for the agent to believe that no other action results in greater expected utility.*<sup>4</sup>

In this formulation, which can be traced back to (Bentham, 1789), an algorithmic system is expected to understand the consequences of its actions. Systems are measured by whether they maximize total overall welfare in their *results*. Once again an agent  $\mathcal{A}$  can be trained, which will make a decision  $\mathbf{d}_i$  for each evidence set  $\mathbf{E}_i$ . But here, you also assign a utility penalty or gain  $\mathbf{u}_i$  for each of those decisions. Rather than judge the algorithm based on whether it followed consistent rules, we instead seek to maximize *overall* gain for all  $N$  decisions that would be made by agent  $\mathcal{A}$  - morality of an agent is equal to  $\sum_i^N \mathbf{u}_i$ .

This is a very different worldview! Here, the burden of provable ethical decision-making no longer lands on transparency in the algorithm or consistency of a classifier over time. Instead, proof of ethical behavior rests on our ability to observe the consequences of the actions the agent takes. One could argue that consequences are hard to estimate and hence we can pick a random action. But that would be irrational. Hence, the principle judges an action by whether the agent acts according to its rational belief of maximizing the expected utility, rather than by the actual consequences. If the agent is wrong then the action turns out to be a poor choice, but nonetheless ethical because it was a rational choice.

In (Crawford, 2017), the author appeals to researchers to actively consider the subgroups that will be harmed or benefited by the automated systems. This plotting of expected consequences and their exhaustive measurement takes precedence in utilitarian ethics, de-prioritizing the interpretability or transparency of the learned model or features that govern our agent. For machine learning researchers, this would mean shifting the focus toward building rich and exhaustive test datasets, cross-validation protocols, and evaluation suites that mirror real-world applications to get a better measurement of impact.

From this work, we might see an initial drop in reported accuracy of our systems as we develop broader test sets that measure the utility of our systems; however, we would then expect overall accuracy on those broad test sets to be the primary measure of ethical fitness of the classifiers themselves. Subgroup-based parity metrics would fall by the wayside in favor of overall accuracy on data that mirrors the real world.

## 6.3 Applying Ethics to NLP systems

We apply the generalization principle in §6.3.1 and §6.3.2 and respect for autonomy in §6.3.3 and §6.3.4.

---

<sup>4</sup>From (Hooker, 2018).

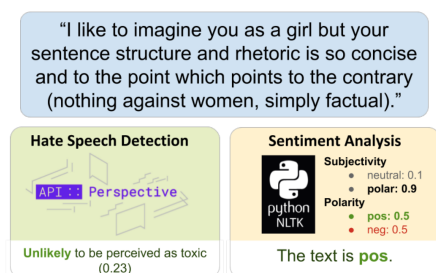


FIGURE 6.1: Micro-aggressive comment and its scores by state-of-the-art hate speech detection and sentiment analysis tools (Breitfeller et al., 2019).

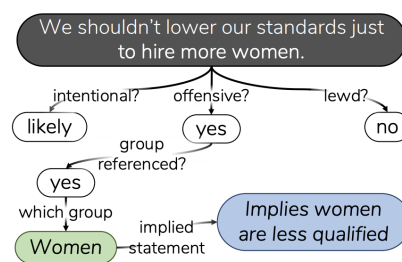


FIGURE 6.2: NLP system flagging the micro-aggressive comment as offensive and generating the reasoning for flagging it (Sap et al., 2020).

FIGURE 6.3: Examples of flagging micro-aggression comments by different NLP systems.

### 6.3.1 Question-Answering Systems

Question-answering (QA) systems have made a huge progress with the recent advances in large pre-trained language models (Devlin et al., 2019; Radford et al.; Guu et al., 2020). Despite these improvements, it is difficult to know how the model reached its prediction. In fact, it has been shown that models often obtain high performance by leveraging statistical irregularities rather than language understanding (Poliak et al., 2018; Geva et al., 2019; Gururangan et al., 2018). The result is that when a QA system is wrong it is difficult for an end user to determine why it was wrong. Presumably, the user would not know the answer to question in the first place, and so it would be difficult to determine even *that* the QA system was wrong.

The act of widely deploying such a QA system is in conflict with the generalization principle. For example, a QA system  $G$  is unsure of its prediction  $\mathcal{A}$  and does not know how it arrived at the answer. Instead of notifying the user about its incapacity to reach the prediction,  $G$  decides to return the prediction  $\mathcal{A}$  due to the following reasons  $\mathcal{R}$ : (1)  $G$  believes that the user does not know the answer and hence (2)  $G$  believes that the user will trust its answer and not ask for reasons on arriving at the prediction. If all QA systems operate like this, users will lose trust in QA systems being able to answer their questions reliably and no longer use them. This contradicts assumption  $\mathcal{R}2$ , violating the generalization principle. This issue goes deeper than a matter of the (in)accuracy of the answer; explainability is still important for a near-perfect QA system. First, the source of an answer could be fallible (even if the content was interpreted correctly), in which case it is important to be able to point which sources were used. Second, answers can often be ambiguous, so a user might naturally ask for clarification to be sure of what the answer means. Finally, it is natural for humans to build trust when working with a system, and explainability is an important step in this process.

Attention weights have been widely used for explaining QA predictions. Attention weights learnt by neural models denote the words or phrases in a sentence that the model focuses on. Hence, words or phrases with high attention weights are considered as explanations to the QA predictions. But these weights do not reliably correlate with model predictions, making them

unsuitable for explainability (Pruthi et al., 2020; Serrano and Smith, 2019; Jain and Wallace, 2019). Recently, generating natural language explanations (Rajani et al., 2019; Laticinnik and Berant, 2020) for predictions has gained traction. These methods train a language generation model to generate explanations for the QA predictions. Using a black-box model for text generation, though, pushes the same problem further down the line. Part of the issue with both of the aforementioned methods is that the “reasoning” for the answer is determined *after* the answer has been generated (i.e., reasoning should inform the answer, not vice-versa).

**The way forward:** A method which reaches the prediction through reasoning would be more in line with the generalization principle. For example, reaching the prediction through traversal of a knowledge graph. This has been used in scenarios where a knowledge base exists (Han et al., 2020; Jansen et al., 2018) for a QA system as well as in dynamic graph generation to reach the prediction (Liu et al., 2020; Rajagopal et al., 2020; Bosselut and Choi, 2019). In these methods, the reasoning is part of the process to generate the final answer, which is more suitable in failing gracefully and building user trust.

### 6.3.2 Detecting Objectionable Content

Social media platforms have made the world smaller. At the same time, the world has seen a surge in hate-speech, offensive language, stereotype and bias on online platforms. These online platforms have traffic in the millions of textual comments, posts, blogs, etc. every day. Identifying such objectionable content by reading each item is intractable. Hence, building an NLP system which can read textual data and flag potential objectionable content is necessary. These systems can reduce the burden on humans by reducing the number of posts that need to be seen by human eyes.

The pivotal role NLP systems play in flagging such content makes the ethical considerations important. Fig. 6.3 shows a microaggressive comment and its scores by a state-of-the-art (1) hate speech detection system and (2) sentiment analysis system. Since these systems rely on surface level words or phrases to detect such (overt) comments, they tend to miss subtle (covert) objectionable content (Breitfeller et al., 2019). If such NLP systems are used universally, then the users of hate speech will discover ways to phrase the same meaning with different words (as illustrated above). Thus, the NLP content flagging system will not be able to detect objectionable content, and there will be no point in deploying it. This contradiction suggests that NLP systems must not make their predictions based only on superficial language features but instead seek to understand the intent and consequences of the text presented to them. Hence, they should generate reasons for flagging posts to facilitate the decision making of the human judges and also to provide evidence about the accuracy of their predictions.

**The way forward:** An example of objectionable content is microaggression (Fig. 6.1). According to Merriam-Webster, microaggression is defined as a “comment or action that subtly

and often unconsciously expresses a prejudiced attitude toward a member of a marginalized group (e.g. racial minority).” Microaggressions are linguistically subtle which makes them difficult to analyze and quantify automatically. Understanding and explaining why an arguably innocuous statement is potentially prejudiced requires reasoning about conversational and commonsense implications with respect to the underlying intent, offensiveness, and power differentials between different social groups. Breitfeller et al. (2019) provide a new typology to better understand the nature of microaggressions and their impact on different social groups. Fig. 6.2 presents such a comment and how we would like the NLP systems to annotate such content. Sap et al. (2020) perform the task of generating the consequences and implications of comments which is a step towards judging content based on its meaning and not simply which words it happens to use. Although such an aim does not automatically solve the problem, attempting to uncover the deeper meaning does not result in an inconsistency or violation of the generalization principle.

### 6.3.3 Machine Translation Systems

Machine Translation (MT) systems have reduced language barriers in this era of globalization. Neural machine translation systems have made huge progress and are being deployed by large companies to interact with humans. But facilitating human-to-human interaction requires more than just simple text-to-text translation, it requires the system to *interpret* the meaning of the language. This requires a greater sensitivity to style, intent, and context on the part of MT systems.

When an MT system acts as an interpreter for a user, it is essentially speaking for the user when conveying the translated message. Speaking for one’s self is within one’s sphere of autonomy, but by using the MT system the user has implicitly consented to it representing the user. That being said, the operating assumption for most users is that the MT system will simply translate the source language into the target language without changing the meaning. Yet on occasion, differences or ambiguities between languages require either contextual knowledge or further clarification on what is being said. If the MT system encounters such ambiguities, the user must be *informed* of such occurrences so that she can *consent* to the message which the system ultimately conveys. Moreover, the user must also be *informed* of the failure cases in the MT system rather than it producing an entirely incorrect translation.

For example, when translating from English to Japanese, there is a mismatch in the granularity of titles or honorifics used to address people. In English, “Ms.” and “Mr.” is an appropriate way to address a school teacher who does not hold a doctorate. On the other hand, in Japanese it would be disrespectful to use the more common “-san” honorific (the rough equivalent of “Ms.” or “Mr.”) in place of “-sensei” which refers specifically to teachers or mentors and shows them a special level of respect. If the MT system cannot reasonably infer how to resolve the ambiguity in such situations, the English speaker should be *informed* about it. The English speaker must

be notified that such an ambiguity needs to be resolved because there is a risk of offending the Japanese speaker otherwise.

In general, there is a trade-off in translation between literality and fluency in certain situations like the translation of idioms. Idioms are especially problematic when considering autonomy because there are multiple strategies to translating them which are not only difficult in and of themselves to execute, but deciding which one to use requires the interpreter (i.e., MT system) to understand the intent of the user. Baker (1992, Ch. 3) identifies five different methods for translating idioms:

1. Using an idiom of similar meaning and form; directly translating the idiom achieves the same effect
2. Using an idiom of similar meaning but dissimilar form; swapping out an equivalent idiom with a different literal meaning
3. Translation by paraphrase; simply explaining the idiom plainly
4. Translation by omission
5. Translation by compensation; for example, omitting idioms in certain locations and adding them in elsewhere to maintain the same overall tone

For example, in casual conversation, an MT system may prefer strategies 1, 2, and 5 to maintain a friendly tone, but in a high-stake business negotiation, it would be more prudent to play it safe with strategy 3. An MT system must be sensitive to the user's intent since choosing an inappropriate translation strategy could violate her autonomy.

While para-linguistic conduct may fill the gaps for in person interaction, if the interaction is happening only via the textual modality, then there is minimal room for such conduct. Additionally, a recent study (Heinisch and Lušicky, 2019) shows that 45% of the participants reported that they expect MT output, in professional and private contexts, to be useable immediately without any further editing. However, post-study, this expectation was not fulfilled. The work further shows that the expectation of the type of errors is also different from the errors in the outputs of the MT system. For example: only 6% of the participants expect that the output would be useless, but after reading the output, 28% thought that the output was useless. The participants in this study had different levels of experience with MT systems (frequent vs. rare users) and used MT systems for different functions (private, professional).

**The way forward:** Mima et al. (1997) drive the early discussion on using information such as context, social role, domain and situation in MT systems. DiMarco and Hirst (1990); DiMarco (1994) advocate for style and intent in translation systems. A study by Hovy et al. (2020) shows that commercial translation systems make users sound older and more male than the original demographics of the users. Recent work (Niu and Carpuat, 2019; Sennrich et al., 2016) has

given specific focus to controlling formality and politeness in translation systems. There is also work directed towards personalizing MT systems (Rabinovich et al., 2017; Michel and Neubig, 2018; Mirkin et al., 2015; Mirkin and Meunier, 2015) while preserving author attributes as well as controlling structural information like voice (Yamagishi et al., 2016). This is a step in the right direction, but we argue that to respect autonomy, translation systems should also obtain explicit informed consent from the user when necessary. Further research must also be done on informing the users about the failure cases of the MT system.

Further research is required in the direction of informing the users about the failure cases of the MT system. For example, in case of ambiguity, textual interfaces can provide multiple suggestions to the addresser along with the implications of using each variant. The user can select the option which best fits their goal. In speech interfaces, the MT system can ask a follow up question to the addresser of the system in case of ambiguity or it can add cautionary phrases to the addressee informing them about the ambiguity. Alternatively, if the system thinks that the input sentence is ambiguous and cannot be translated with reasonable confidence then it can say “I am unable to translate the sentence in its current form. Can you please rephrase it?”. An example scenario where such clarification might be needed is: while translating from English to Hindi if the sentence refers to one’s “aunt,” the MT system should ask a follow up question about maternal vs paternal aunt since they have two different words in Hindi language.

#### 6.3.4 Dialogue Systems

We can find a nuanced application of the autonomy principle in the way that dialogue systems, especially smart toys or virtual assistants like Alexa and Google Home, interact with children.

One expression of a parent’s autonomy<sup>5</sup> is generally in deciding whom their child may interact with. For example a parent would permit interaction with a teacher but not a random stranger. In the case of a parent purchasing and using a virtual assistant at home, they are implicitly *consenting* to their children interacting with the assistant, and the issue arises from the fact that they may not be *informed* as to what this interaction entails. To an adult, a virtual assistant or dialogue-capable toy may seem like just another computer, but a 7-year-old child might view it as “more capable of feelings and giving answers”—a step in the direction of assigning personhood (Druga et al., 2017). Furthermore, while humans have had thousands of years to learn about human-human interaction, we have only had a half-century to learn about the effects of human-machine (and thus, child-machine) interaction (Reeves and Nass, 1996).

We suggest two key areas which are important for dialogue system researchers: (1) they must answer the question of what unique social role do dialogue systems fulfill—that is, in what respects can they be regarded as human-like vs. machine-like, and (2) the dialogue systems must have some way of modeling the social dynamics and cues of the interlocutor to fulfill the social role properly.

---

<sup>5</sup>This is technically *heteronomy*, but this examples comports with the spirit of *respect for autonomy*.

**The way forward:** There is a fair amount of research on the social aspects of human-computer dialogue both in general and specifically with regards to children (Druga et al., 2017; Shen, 2015; Kahn Jr et al., 2013). Although it is difficult to gain a complete understanding of how dialogue systems affect the development of children, the most salient facts (e.g., children regarding virtual assistants as person-like) should be communicated to parents explicitly as part of parental controls. We advocate for a “kids mode” to be included with these virtual AI assistants which would provide the feature of *parental control* in accordance with respect for autonomy. This mode would be aware that it is talking to children and respond accordingly. NLP can also help in selecting content and style appropriate for children in these AI agents. Additionally, parents can be provided with fine-grained control over the topics, sources and language that would be generated by the agent. For example, the parent can select for a polite language and topics related to science to support their child’s development efforts. Much research has focused on controlling topics (Kim et al., 2015; Jokinen et al., 1998), style (Niu and Bansal, 2018b), content (Zhou et al., 2018; Zhao et al., 2020a; Dinan et al., 2018) and persona (Zhang et al., 2018) of dialogue agents which can be used for this purpose.

## 6.4 Ethical Decision Making with NLP

So far we have discussed how NLP systems can be evaluated using ethical frameworks and how decisions made by such systems can be assisted by these theories. NLP can also aid in making decisions in accordance with the deontological framework. Recall that the generalization principle judges the ethical standing of pairs of actions and reasons; these pairs could be extracted with various NLP techniques from textual content. In the case of flagging objectionable content (§6.3.2), extracting the deeper intents and implications corresponds to the reasons for the action of flagging the content. Another example is building an automatic institutional dialog act annotator for traffic police conversations (Prabhakaran et al., 2018). These dialog acts contain the rationales of the two agents in the conversation: the police officer and the civilian stopped for breaking traffic rules. The decision made by the police officer (the action) can then be judged to be in accordance (or not) with a human-selected set of ethically acceptable action and rationale pairs. Similarly, for court hearing transcripts, the rationales of the arguments can be extracted and the verdict of the judge can be checked using them (Branting et al., 2020; Aletras et al., 2019). NLP tools such as commonsense knowledge graph generation (Bosselut et al., 2019a; Saito et al., 2018; Malaviya et al., 2019), semantic role labeling (Gildea and Jurafsky, 2000), open domain information extraction (Angeli and Manning, 2013) etc., can be used to extract rationales, entities from text and also find relations between them to better understand the underlying intent of the text.

## 6.5 Discussion

We provide a broad discussion on the limitations of the principles chosen in this work and the issue of meta-ethics. Moreover, we emphasize that ethical research is not merely a checklist to be satisfied by abiding to the principles mentioned here. It requires our persistent attention and open-minded engagement with the problem.

One limitation of this work is in the principles that we choose.<sup>6</sup> For example, the interaction of machine learning and privacy is of huge ethical importance. While the respect for autonomy may address this issue in part, it would be more productive to utilize a deontological principle to the effect of the *right to privacy* with which such matters can be judged.

Another instance is that in this work, we have not discussed the principle of *interactional fairness* (Bies, 2015, 2001) which refers to the quality of interpersonal treatment including respect, dignity, and politeness. With the increasing amount of interaction between humans and machine, the natural language generation systems can be evaluated with this principle. Systems which show respect and dignity to users as well as generate polite language can enhance the degree of interactional justice, which can in turn enhance utility (e.g., trust, satisfaction). Our work on politeness transfer (Madaan et al., 2020b) which aims at generating polite language would be rated higher on interactional fairness compared to system which don't generate polite language.

Additionally, there are broader limitations in using deontology as our ethical framework. In scenarios where there are no *a priori* duties or rights, taking a consequentialist approach and optimizing the effects of ethical guidelines could be more felicitous. For example, the specific rights and duties of autonomous AI systems is not immediately clear. Thus, determining ethical recommendations based on what leads to the most responsible use of the technology would be clearer than selecting appropriate rights and duties directly. Furthermore, rule-based formulations of consequentialism make ethical judgments based on rules, where the rules are selected based on the consequences. Such theories combines some of the benefits of both deontology and consequentialism.

The above difficulties are part of the larger issue of metaethics, that is, the discussion and debate on how to choose among different ethical theories. Within deontology, there is no one standard set of rules. And even within the generalization principle, there is considerable leeway to what “conceivable world” or “logically consistent” mean and how they could be applied to decision making. While presenting a universally accepted ethical theory is likely impossible, metaethical considerations can still be relevant, especially in light of the application of ethical theories. As the field of NLP gets more accustomed with theories of ethics, it will be fruitful to compare the strengths and weaknesses of different ethical theories within the context of NLP and machine learning.

---

<sup>6</sup>Kant would argue that the generalization principle can account for all ethical decisions, but we make no such claim.



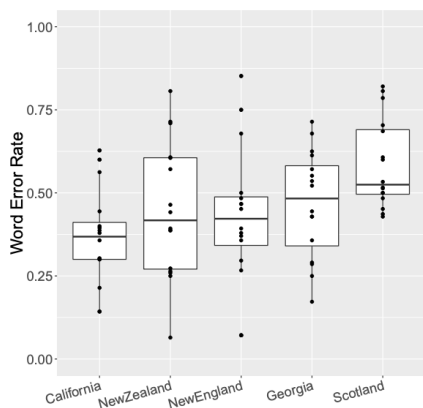


FIGURE 6.4: YouTube automatic caption word error rate by speaker's dialect region. Points indicate individual speakers.

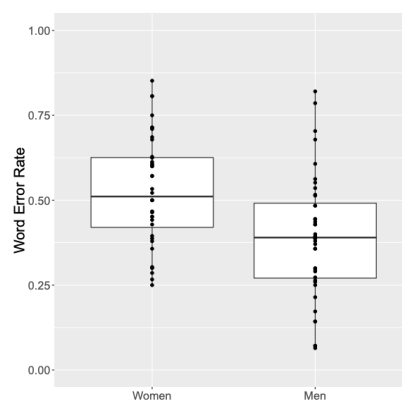


FIGURE 6.5: YouTube automatic caption word error rate by speaker's gender. Points indicate individual speakers.

FIGURE 6.6: Word Error rate plots for gender and dialect (Tatman, 2017)

## Real World Scenarios

These philosophical frameworks do not always diverge in their evaluation of models. Sometimes, models have unambiguously unethical gaps in performance. The exploration from (Tatman, 2017), for instance, shows the difference in accuracy of YouTube's automatic captioning system across both gender and dialect with lower accuracy for women and speakers from Scotland (shown in Figure 6.6, reproduced from the original work). This study shows how this system violates the utilitarian principle by negatively impacting the utility of automatic speech recognition for women and speakers from Scotland. YouTube's model also violates the generalization principle, by incorporating superfluous information about speakers in the representation space of the models. The authors suggest paths forward for improving those models and show that there is room to improve.

But sometimes, solutions highlight differences across ethical frameworks. In (Hovy, 2015), for instance, the author shows that text classification tasks, both sentiment and topic classification, benefit from embeddings that include demographic information (age and gender). Here, the two ethical frameworks we have discussed diverge in their analysis. The generalization principle would reject this approach: age and gender shouldn't intrinsically be used as part of a demographic-agnostic topic classification task, if the number of sources of information is to be minimized. Similarly, changing the feature space depending on the author, rather than the content of the author's text, does not result in models that will make the same decision about a text independent of the identity of the author. The utilitarian principle, in contrast, aligns with the Hovy approach. A more accurate system benefits more people; incorporating information about authors improves accuracy, and so including that information at training and prediction time increases the expected utility of the model, even if different authors may receive different predictions when submitting identical texts.

For an alternate example in which the generalization principle was prioritized over utility, consider the widely-cited coreference resolution system of (Bolukbasi et al., 2016). This paper found that word embeddings used for coreference resolution were incorporating extraneous information about gender - for instance, that doctors were more likely to be men, while nurses were more likely to be women. This and similar work in “debiasing” word embeddings follows the generalization principle, arguing that removing information from the embedding space is ethically the correct action, even at the expense of model accuracy. This work finds that it can minimize the drop in expected utility, reducing F1 scores by less than 2.0 while removing stereotypes from their model. However, in a fully utilitarian ethical framework, even this drop would be unjustifiable if the model simply reflected the state of the world, and removing information led to reduced performance.

## 6.6 Conclusion

Two principles of deontological ethics—namely the *generalization principle* and *respect for autonomy* via *informed consent*—can be used to decide if a decision was ethical. Despite the limitations of these principles, they can provide useful insights into making NLP systems ethical. Through the four case studies discussed (§6.3), it is demonstrated how these principles can be used to evaluate the decisions made by NLP systems and to identify the missing aspects. For each of the case studies, potential directions for NLP research are presented to move forward and make the system ethical.

We further provide a summary on how NLP tools can be used to extract reasons and rationales from textual data which can potentially aid deontological decision making. Note that we do not advocate deontological ethics as the only framework to consider. On the contrary, we present this work as the first of its kind to illustrate *why* and *how* ethical frameworks should be used to evaluate NLP systems. With this work, we hope the readers start thinking in two directions: (1) using different ethical frameworks and applying the principles to NLP systems (like the case studies in §6.3), and (2) exploring the directions mentioned in the case studies of this paper to improve current NLP systems.

**Controllable Text Generation for ethical impact:** Controllable text generation plays a pivotal role in making these systems ethical. It can be used to ensure transparency in model predictions, personalizing models, responsible and faithful prediction toward various demographic groups and generating consequences and implications. As discussed in the case studies (§6.3), the following controllable text generation solutions need to be explored: (1) dynamic graph generation for QA predictions, (2) generating the consequences and implications of comments is useful in detecting microaggression. It is a step towards judging content based on its meaning and not simply which words it happens to use. (3) controlling formality, politeness as well as personalizing translation systems by controlling author attributes, structural information and voice. (4) controlling topics, style, content and persona of dialogue agents.

# Chapter 7

## Conclusions

### 7.1 Summary of Contributions

The aim of this thesis was to control for style, content and structure in generation for producing human-like generation. Another important goal of the thesis was to understand the ethical implications of controllable text generation. To fulfill these goals, it was first important to understand the space of controllable text generation, the tasks involved and the challenges it entails. Hence, in chapter 2, I introduce a new schema for controllable text generation which contains five modules - (1) external input, (2) sequential input, (3) generator operations, (4) output, and (5) training objective. The schema organizes prior work and provides an insight into the contributions of the various modules for controllable text generation.

I dedicate one chapter each for controlling style, content and structure in text generation and one for understanding the ethical considerations. In each of these chapters, I detail the specific tasks used controlling each aspect, the datasets used, the modeling techniques, the experiments and results.

For controlling style, I focus on the task of style transfer in chapter 3. Style transfer is the task of rephrasing the text to contain specific stylistic properties without changing the intent or affect within the context. I introduce two new tasks for exploring style transfer: (1) political slant transfer, and (2) politeness transfer. I provide datasets for both of them for further exploration. I develop two novel approaches to perform style transfer with non-parallel data: (1) back-translation for style transfer, and (2) the state-of-the-art tag and generate approach. The details of the code for implementing both the techniques and the trained models is provided in §A. Different types of human evaluations are presented for this task to measure the efficacy of style transfer models along the three dimensions of: (1) style transfer accuracy, (2) preservation of meaning and (3) fluency of generation. Human evaluations are necessary but not sufficient in proving the success of the models for the style transfer tasks. I discuss the limitations of automated metrics in this chapter: (1) classifiers used to measure style transfer accuracy are

not robust and can be easily fooled by polarizing keywords, and (2) BLEU, METEOR and Rouge-L rely heavily on N-gram word overlap and hence cannot effectively measure preservation of meaning. Hence, good metrics to measure the success of style transfer task remains an open problem.

For controlling content, I establish the task of document grounded generation in chapter 4. Document grounded generation is the task of generating text that is coherent to a context and contains information from a document. I introduce two new tasks under the framework of document grounded generation in two different domains: (1) Wikipedia Update Generation and (2) Document Grounded Dialogue Generation. For both the tasks, I release large datasets for further investigation. I develop two extensions to pre-trained encoder-decoder models specifically for document grounded generation: (1) Context Driven Representation model, and (2) Document Headed Attention model. The dataset and the code for implementing the proposed models is provided in §B. Automated metrics of BLEU, METEOR and Rouge-L are used to evaluate the success of the models. These metrics don't measure the notion of grounding. I develop two new human evaluations specifically for the document grounded generation task: (1) *closeness*, measures how close to the reference is the generated sentence and (2) *relevance*, measures if the generated sentence contains information from the document and is coherent to the context. I perform extensive manual inspection and present an error analysis of the quality of generations in this chapter (§4.3.3). I find that the automatic evaluation and the human evaluations don't give a complete picture of the nature of generations. The automated metrics penalize correct generations with low word overlap. Moreover, I find that especially in the task of document grounded dialogue response generation, more than one response could be correct. But the automated metrics only rely on the sole reference to measure the quality of the generation and this is insufficient. Hence, evaluating document "grounding" and its extent in the generation is hard and an unsolved problem.

For controlling the structure, I focus on the subtask of sentence ordering in chapter 5. I propose a new framing of sentence ordering task as a constraint solving task and use topological sort for this task. I introduce a human evaluation for the sentence ordering task. Additionally, I provide the details of the code for implementing topological sort and constraint learning in §C.

I finally discuss the ethical implications of controllable text generation in chapter 6. Ethics can also be used to evaluate NLG systems. I present an overview of two deontological principles (*generalization principle* and *respect for autonomy*) along with a discussion on their limitations with a special focus on NLP. I illustrate four specific case studies of NLP systems which have ethical implications under the deontological principles and provide practical directions to alleviate these issues. The analysis presented in this chapter suggests that controllable text generation techniques can be used in many cases to make NLP systems ethical.

## 7.2 Future Directions

This thesis has focused on controlling the generation process to build socially aware, content-rich and ethical NLG systems. This section is split into two parts: (1) Section §7.2.1 discusses the broad directions for future work and (2) Sections §7.2.2, §7.2.3, §7.2.4 and §7.2.5 identifies and discusses a few practical projects which can take the field in the directions mentioned in §7.2.1.

### 7.2.1 Broad Directions for Future Work

**Expand attributes:** This thesis has explored the tasks of style transfer, document grounded generation and sentence ordering. Future work can focus on controlling more broad attributes like controlling persuasion and bias in generation. Recent work has focused on generating language that is not biased (Pryzant et al., 2020; Ma et al., 2020). The techniques discussed in this thesis can be explored to solve these tasks. Persuasive language has been analyzed in debate and argumentation setting (Atkinson et al., 2019; Luu et al., 2019; Tan et al., 2016) as well as persuasive strategies have been studied in dialogue setting (Wang et al., 2019c; Tian et al., 2020). So far, little work is done to use it as an attribute to control the generation process. Future work can aim at controlling persuasion in text generation for these tasks.

**Combine attributes:** Most real world applications demand the control of multiple attributes for seamless interactions with humans. For example, an FAQ both has to make sure that the content generated in the response is accurate as well as tune the level of detail in the response (style) according to the expertise of the user. This thesis has studied controlling each attribute (style, content and structure) individually. Future work can focus on controlling multiple attributes simultaneously like controlling both style and content at the same time. Some recent work has focused on controlling multiple attributes in dialogue response generation (See et al., 2019) and story generation tasks (Hu et al., 2020a). Future work can expand on these ideas and aim to control various aspects of controllable text generation tasks simultaneously.

**Multilingual Controllable Text Generation:** Finally all the tasks explored in this thesis are in English language and controlling the generation in multiple languages is a crucial future work. The current multilingual work primarily focuses on classification tasks (Vulić et al.; Conneau et al., 2018) like semantic similarity and natural language inference, structured prediction (Nivre et al., 2016; Pan et al., 2017) tasks like part of speech tagging and named entity recognition, Question Answering task (Lewis et al., 2020) and retrieval tasks (Hu et al., 2020b). Generation tasks such as multilingual summarization (Scialom et al., 2020), image caption generation (Jaffe, 2017), generating multilingual text from RDF data (Gardent et al., 2017) and code switching (Parekh et al., 2020) are recently gaining attention. Future work can target designing

the following multilingual tasks and work on collecting datasets, designing new models and evaluating such tasks:

- Multilingual document grounded generation: given a document in one language, generate document grounded pieces of text in multiple languages from the same source document.
- Multilingual dialogue response generation conditioned on the given persona, topic, document etc;
- Multilingual story generation grounded in the same plot line.
- Sentence ordering task for multiple languages.

**Evaluation:** This thesis also brings out the limitations of the current automated metrics for evaluating generation tasks (§3.3, §4.3.3). Hence, this thesis has developed new human evaluations for document grounded generation (§4.3.2) and sentence ordering (§5.2.3). But human evaluation studies are expensive and are generally not reproducible. Hence, there is a need of better automatic evaluations measures which could be cheap and reproducible. This thesis provides a new dimension of ethics to evaluate NLP systems and shows how ethical principles can be used to evaluate them. Yet, there remains a gap of evaluating multiple aspects of NLG tasks through automated ways. For example, currently there are no good automated metrics for meaning preservation in style transfer and for evaluating grounding in document grounded generation. Hence, building good automated metrics or tools for evaluating the various tasks of generations remains a strong line of research for future work.

## 7.2.2 Exploring Controllable Text Generation Techniques

The future work can focus on empirical evaluation of the proposed schema. Such an evaluation can select a few controllable text generation tasks and provide empirical insight into which of the described techniques work better or worse for different tasks. The techniques described for each of the five modules can be experimented under similar settings to gain understanding of the contribution of each of the modules in controlled text generation. It is also possible that a combination of techniques is suitable for some tasks. Such experiments can provide new directions to explore for controllable text generation.

## 7.2.3 Style Transfer

**Building Effective Style Representations:** Pre-trained language models like BERT have shown to contain syntax information (Li and Eisner, 2019; Hewitt and Manning, 2019) and relational knowledge (Petroni et al., 2019). These ideas can be extended to get a stylistic representation from BERT representation. Building effective style representation could be used in

cross domain classification tasks. For example, if data on a style  $S_1$  is present in abundance in domain  $T_1$  but there is limited availability of data in another domain  $T_2$ . Then, a good representation of  $S_1$  can be built by leveraging the data in  $T_1$  and it can be used for  $T_2$ .

We have performed preliminary analysis by simple averaging of BERT representations to get the style representation. Details are provided in Appendix §A. We acknowledge that the simple averaging technique will not be effective for all types of styles, especially it may not be effective for highly content coupled styles. Hence, future work can extend this direction and design better ways of extracting style borrowing ideas from (Kumar et al., 2019; Li et al., 2015).

**Understanding Style:** Understanding the complexity of style, can help in designing suitable methods for style transfer. Hence, future work can focus on building a computational approach towards understanding the different complexities of style. For example, to determine whether a style is content coupled or not, a classifier which only uses bag of words can be used. Other features such as POS tag sequences, sentence structure, usage of function words etc may contribute in defining a particular style. Based on this understanding, we can formulate different types of transformations that are possible for that style. For example, for content decoupled style like politeness, a simple delete, replace or add pipeline would suffice to perform style transfer. Contrastingly, for heavily content coupled styles like sarcasm which depends on the context, this simple transformation may not work. The specific models would be dependent on the type of transformations that need to be performed.

Future work can focus in the two directions of *lexical understanding* and *structural understanding* of the styles. The lexical understanding will give an understanding of how content dependent is the style. This includes, understanding the N-grams features, the distribution of function and content words in each style. This would give a better understanding on the type of words the style is dependent on. Style may not necessarily be only defined by surface level features. Some styles are dependent heavily on the underlying structure of the sentence. Structural features include POS tag N-gram features, features from parse trees, as well as additional features such as depth and breadth of the trees. Some preliminary experiments are shown in Appendix §A which can be studied in depth in the future.

**Multiple Style Transfer:** This thesis primarily studies transferring one style at a given time. Some works have explored the idea of generating text by controlling multiple target styles (Lample et al., 2018; Logeswaran et al., 2018a). Prior work has not been successful in controlling all the styles simultaneously with high accuracy. This could be due to the varying complexities of the styles. Also, it is important to understand the correlations between styles for simultaneous multiple style transfer task. For example if style  $s_1$  and  $s_2$  are highly correlated but if we are only trying to change one and the other constant then that might be a difficult task. Hence, understanding their styles and their interdependence is important for this task. Future work can focus on formalizing the task controlling multiple styles simultaneously, curating good datasets for the task and building effective models to solve it. A two part

evaluation is required: (1) how successfully is the model transferring each of the target styles in question, and (2) when the model is only supposed to transfer two (say) of the three target styles, then is the third style actually preserved.

An interesting domain to observe multiple style transfer is in case of demographic attributes like gender, age, education etc. Kang et al. (2019) provides the PASTEL dataset which contains parallel dataset for different demographic attributes.

**Degree of Style Transfer:** This thesis as well as most of the contemporary work treats style as a discrete variable. It focuses on the task of completely transferring from one polarity to the other. Some real world styles may require a fine-grained control and a transfer mechanism where you can control the degree of the style. An example is Yelp reviews, where the review corresponding to the number of stars would be different. Most sentiment modification tasks, convert this into a binary problem of transferring from negative to positive sentiment and vice versa. But in case of Yelp reviews, you may want more degrees of style transfer. The task then entails transferring a review with one star to a review with two or three stars instead of five stars. Wang et al. (2019a) presents a preliminary study of controlling the degree in sentiment modification task. Future work can aim to extend this idea for more styles and work towards building datasets and models to explore this direction.

#### 7.2.4 Document Grounded Generation

**Evaluation of the task:** The limitations of the automated metrics currently used are discussed in §4.3.3. The current metrics do not evaluate the generative models for how much content is transferred from the external source of document to the generated output. This is a crucial dimension along which the document grounded generation models should be evaluated. In (Prabhumoye et al., 2019b), we propose two dimensions along which the grounded generation is evaluated: *close to reference* and *coherent to context*. Close to reference measures how close the generated output is to the reference text. Coherent to context measures whether the generated text contains key information from the external document and fits the input context. This has been measured by human judges using A/B testing (Bennett, 2005) comparing two models. Future work can focus on building an automatic metric for these two evaluations.

Potential directions include using information extraction systems like OpenIE (Angeli et al., 2015), frames based on intent, effect and reaction (Bosselut et al., 2019b), or keyword extraction system (RAKE) (Rose et al., 2010) to extract information. Let the information extracted from the reference be  $\mathbf{i}_r$ , the input text be  $\mathbf{i}_s$ , the external document be  $\mathbf{i}_d$  and the generated text be  $\mathbf{i}_g$ . For the *close to reference* metric,  $\mathbf{i}_r$  can be compared with  $\mathbf{i}_g$  using different metrics like Jaccard similarity score or BLEU. Similar metrics can be used to compare  $\mathbf{i}_g$  with  $\mathbf{i}_d$  and  $\mathbf{i}_s$  for the *coherent to context* metric. Alternatively, cosine similarity can be calculated between the BERT representations of  $\mathbf{i}_r$  and  $\mathbf{i}_g$ , and  $\mathbf{i}_g$  with  $\mathbf{i}_d$  and  $\mathbf{i}_s$ .



The choice of information extraction system and the comparison metric can be based on what is considered as *information* for the task in question.

**More tasks:** I focus on two concrete tasks in this thesis: Wikipedia Update generation and document grounded dialogue response generation. Future work can explore more concrete tasks that fit the framework of document grounded generation and apply the techniques described in this thesis.

The update summarization task (Dang and Owczarzak, 2008) can use the modeling insights from this thesis. This task encompasses updating an existing summary with new information from a batch of documents. In case of repeated updates of the same summary with new documents, one can apply the techniques described in this thesis iteratively to the summary.

Reasoning about events and tracking their influences is fundamental to understanding processes. Recently text generation has been used to generate event influences conditioned on a source event and a context (Madaan et al., 2020a). This work particularly focuses on generating an event given a source event and a scientific procedural text about the events. Currently, only large scale pre-trained language models have been explored to solve this task. This task can fit into the framework of document grounded generation where the scientific procedural text can be considered as the document and the source event can be considered as the context. Hence, future work can aim to use the techniques described in chapter 4 for solving this task.

Many tasks in financial sector involve generating reports and summaries which can be fit the document grounded generation framework (described in §7.3.1).

### 7.2.5 Ethical Considerations

**Generating balanced datasets:** Downstream tasks are influenced by the skew in training sets like the sentiment analysis task is affected by the gender confound (Hovy et al., 2015) and the part of speech (POS) tagging task is affected by the age confound (Hovy and Søgaard, 2015). Webster et al. (2018) release a gender balanced dataset for coreference resolution task. Zhao et al. (2017) also explore balancing dataset with gender confound for multi-label object classification and visual semantic role labeling tasks. Data augmentation by controlling gender attribute is an effective technique in mitigating gender bias in NLP processes (Sun et al., 2019; Dinan et al., 2020a). Wei and Zou (2019) explore data augmentation techniques that improve performance on various text classification tasks. Controllable text generation techniques can prove to be useful in augmenting training data by generating data representative of minority class label. Using the techniques described in 3, it is possible to generate data by controlling for the gender attribute (style) while preserving the content. Future work can aim at producing demographically balanced datasets for the above mentioned NLP tasks using controllable text generation.

## 7.3 Broader Impact

So far we have talked about the technical improvements in this thesis in terms of scientific contributions like new tasks, datasets and practical tools. I would like to conclude my thesis by discussing the merits and contributions of my thesis outside the scientific research community.

### 7.3.1 Impact beyond NLP

As discussed earlier, the work done in this thesis has applications in building dialogue systems of various types like personal assistants, FAQ bots, empathy bots, automated report and story generation as well as in writing assistant tools. Within the scope of NLP applications controllable text generation techniques could be used in any system that gives information to humans, to control demographic attribute in machine translation systems, and to control structure and content in summarization. But controlling text generation has applications outside of the NLP field. Below, I outline a few applications outside NLP in three field of *education*, *healthcare* and *finance sector*.

**Education:** NLP tools are already perforating in the education domain to assess essays written by students (Mayfield and Black, 2020; Mayfield et al., 2019), to identify students who might dropout from MOOCs (Yang et al., 2013; Rosé et al., 2014; Wang et al., 2015) and to analyze student collaborations (Rosé et al., 2008; Kumar et al., 2007). Controllable text generation can be used to build automated tutoring bots. Furthermore it can be used to build personalized tutoring bots. The instructions provided by teachers in classrooms cater to the whole classroom. Some students may need further explanation or more examples to reinforce the concepts. Such detail can be provided by personalized tutoring bots. Duolingo application (Duolingo, 2011) which is a system to learn new languages already offers personalized learning experience. The exercises are tailored to the learning curve of individuals. They also generate paraphrases of the translations that language learners are likely to produce to augment their learning abilities (Mayhew et al., 2020). These tasks of generating personalized examples and paraphrases of translation according to expertise level can hugely benefit from controllable text generation techniques. Another popular application that can benefit from controllable text generation is Grammarly. Grammarly (Grammarly, 2009) is a writing assistance tool which also provides stylistic and tone recommendations in writing. Writing assistance agents are also embedded in Microsoft Word which provide stylistic suggestions such as formal language, conciseness and clarity (Microsoft, 2021). These writing assistance tools can also be added to tutoring bots to support the learning of students. Such bots could ideally help spread education in poorer nations or underprivileged groups of people. But I completely acknowledge that such technology may not be accessible or affordable by these groups.

**Healthcare:** NLP has proven to be essential even in the healthcare domain. Most of the current work focuses on medical text by identifying medical terminologies and ontologies using tagging, parsing, entity recognition and coreference resolution techniques (Bhatia et al., 2020; Rumshisky et al., 2020). Information extraction from clinical text or medical conversations has also gained traction in recent years (Ding et al., 2020; Poug   Biyong et al., 2020). The pre-trained language models have not proven to be as useful in this domain due to usage of specific terminologies that will not be found outside of this field (Alsentzer et al., 2019; Huang et al., 2019a). Little work is done in generating medical reports from patient-doctor conversations (Enarvi et al., 2020) or building healthcare bots that can assist patients in providing them with relevant information and setting reminders for intake of prescriptions. Building agents that can assist in mental healthcare is also gaining attention. All these applications can be improved using controllable text generation techniques that can help generate accurate reports by grounding in the content of conversations or generate responses tailored to the level of detail required by the patients. Crisis Text Line (Line, 2013) is a not-for-profit organization that provides mental health support via text service like SMS. Zhang and Danescu-Niculescu-Mizil (2020) study the strategies and objectives in counseling conversations using the data provided by Crisis Text Line. Controllable text generation could be used to provide recommendations of responses conditioned on these strategies and objectives.

**Finance:** In the finance sector “*time is money*” and it is crucial to handle the growing number of insights that are being produced by ever-increasing data through automated forms of analysis. Controllable text generation can be used to make sense of the large amount structured as well as non-structured data in the finance industry. Document grounded generation can help in varied number of tasks in this space such as automatically generate financial reports and executive summaries from huge amounts of unstructured data in the form of company documents. Another task could be to generate risk and underwriting reports grounded in analysis of client profile documents. Notably, an important future direction could be used to generate short summaries from various news sources to understand the overall sentiment and public perception of world events or organizations. In addition, it can be used generate suspicious activity reports (SAR) from activity alerts and financial transaction. These efforts would reduce the time spent by analysts on repetitive tasks and hence cut down the cost. Bloomberg already offers automated tools for some of these tasks (Bloomberg, 1981). Some of these tools can be assisted by controllable text generation techniques for faster and cheaper solutions.

### 7.3.2 Impact on Society

The most crucial impact of this thesis is on society; on the lives of the people who would use the applications built by this technology. Hence, it is important to assess if the techniques described in this thesis can help in treating everyone equally or would it marginalize certain groups of people; would it make applications accessible to people of all sections of the society or would it exclude certain groups (Benjamin, 2019). In reality, this technology could have

both benevolent as well as harmful impact on society. I sketch a few scenarios describing both ahead.

The current machines that interact with humans assume a “*One size fits all*” philosophy while giving information to people. Controllable text generation can help in catering to different people and personalizing the communication they receive from machines. Personalization is possible as you can control the style of the generated text using style transfer techniques (§3). Pre-trained language models and pre-trained encoder-decoder models have become pervasive in the scientific research community as well as in applications that interact with humans.

Pre-trained language models (LMs) face the following well-documented issues (Bender et al., 2021): (1) lack of diversity in training data, (2) static data but changing social views, and (3) encoded bias. Although, controllable text generation techniques may not completely solve all of these issues, they can certainly be explored to alleviate them. As described in §7.2.5, it is possible to generate balanced datasets for minority class of data. Style transfer techniques can be explored to generate data from smaller dataset with limited annotations (Zhao et al., 2017; Sun et al., 2019). Once trained, LMs run the risk of ‘value-lock’, where the LM-reliant technology reifies older, less-inclusive understandings. One expensive way of mitigating this would be to retrain LMs with societal changes. But if controllable text generation is used on top of LMs to fine-tune them, then it could be a cheaper way of ensuring the inclusion of new norms, language and ways of communication. Finally, the encoded bias in LMs is well-documented (Basta et al., 2019; Kurita et al., 2019; Zhao et al., 2019). Despite the biased representations provided by LMs, if we can have a stronger control on the content and style of generation then it might be possible to lessen the bias in generated text. I am not intimating that controlling style and content would solve all the current problems in pre-trained LMs but I am asserting that these techniques can be explored and may lead to a path forward in alleviating them.

The broader impact of controllable text generation can also be explored by understanding the way humans interact with technology and how it has the capacity to change human attitudes and beliefs. Kaufman and Libby (2012) explore the concept of experience-taking in changing human behavior and beliefs. Experience-taking is defined as the imaginative process of spontaneously assuming the identity of a character in a narrative and simulating the character’s thoughts, emotions, behaviors, goals and traits as if they were one’s own. When experience-taking occurs, readers adopt the character’s mindset and perspective as the story progresses rather than orienting themselves as an observer or evaluator of the character. Six studies in this work (Kaufman and Libby, 2012) investigated the features of narratives that cause individuals to engage in experience-taking without instruction. Additionally, they investigated how the merger between self and other during experience-taking produces changes in self-judgments, attitudes, and behavior that align with the character’s. These studies find that greater ability of a narrative to evoke experience-taking increases the ability of a reader to simulate the subjective experience of a character which in turn increases the potential that story has to change the reader’s self-concept, attitudes, and behavior. The study found that a first-person narrative

depicting an ingroup character elicited the highest levels of experience-taking and produced the greatest change in participants' behavior, compared with versions of the narrative written in 3rd-person voice and/or depicting an outgroup protagonist. The studies demonstrated that whereas revealing a character's outgroup membership as a homosexual or African American early in a narrative inhibited experience-taking, delaying the revelation of the character's outgroup identity until later in the story produced higher levels of experience-taking, lower levels of stereotype application in participants' evaluation of the character, and more favorable attitudes toward the character's group. Controllable text generation can open new directions for using these studies as firm basis to change human attitudes towards stereotypes on minority groups. For example, controllable text generation could be used to generate powerful narratives in first person which encourage experience-taking and reveal the identity of the group much later in the interaction. Similarly, (Seering et al., 2019) explores a chatbot's social role and how it can be used to maintain moderate growing online communities. This work identifies seven categories of chatbots for this role. This work can be extended to more personas of chatbots which can further be used to change human perspectives on minority groups.

So far we have discussed the positive impacts of this thesis. Like every technological tool, this too can be used for negative applications. Controlling content can give rise to generating fake news or fake text. As described in §4.4, this thesis does not focus on fact checking but tries to ensure that the generation is faithful to the content in the document. Hence, if you provide factually incorrect information to the model, then it will generate inaccurate or fake information. Furthermore, controlling for style can enable such applications to personalize fake news and generate it in a manner that is appealing to each individual (Zellers et al., 2020; Schuster et al., 2020). Controllable text generation could also be used to generate microaggressive comments and hate speech.

Another harmful application of controllable text generation could be its use for propaganda. Propaganda typically requires subtle manipulation of language. Such fine grained control on the generated text can be provided by controllable text generation techniques as the area moves forward. Propaganda often uses strategies such as demonizing the enemy, flooding with misinformation, framing the ideas using stylized lexicons that can potentially sway public sentiment, political ideas and morality (Field et al., 2018; King et al., 2017; Black, 2018). Some of these attributes can be controlled in the generation process to fabricate text that is appealing to certain sections of the society.

The 2016 presidential election of the United States of America was a witness to some of these propaganda tactics. Performance optimizing algorithms were allegedly used to micro-target the audience via automated generation of ads (Lewis and Hilder, 2018). If these propaganda strategies use controllable text generation to automate and control for various attributes such as age, location, gender, race and socio-economic status then we would see a catastrophic ingestion of fake news and misinformation online.

Finally, it would be foolish to assume that technology can solve all the problems of the society. Progress in well-intentioned technology and its integration in existing systems is necessary, but

inevitably it may not solve the deep-seated and complex issues of the society (Morozov, 2013). The solutions mentioned above can only work if researchers invest time and effort to curate suitable datasets and carve the right tasks. Human intervention and thought is required at every stage of the machine learning design lifecycle to prioritize equity and stakeholders from marginalized groups (Costanza-Chock, 2020). Researchers have to be mindful of the entire research design: datasets they choose, the annotation schemes or labeling procedures they follow, the manner in which they decide to represent the data, the algorithms they choose for the task and how they evaluate the automated systems. Researchers need to be aware of the real-world applications of their work and consciously decide to choose to help marginalized communities via technology (Asad et al., 2019). The omnipresence of NLP technologies in society has empowered researchers to bring about change; let's use it to empower others.

# Appendix A

## Appendix for Style Transfer

This appendix details the hyper-parameters of the models described in Chapter 3 and presents examples of the generated sentences for each of the three tasks for the BST and CAE models. It also presents additional experiments with auto-encoders.

### A.1 Details of Training

**Implementation details for the Back-translation (BST) model:** In all the experiments, the generator and the encoders are a two-layer bidirectional LSTM with an input size of 300 and the hidden dimension of 500. The generator samples a sentence of maximum length 50. All the generators use global attention vectors of size 500. These are especially used during the test time to replace the ‘unk’ token. The CNN classifier is trained with 100 filters of size 5, with max-pooling. The input to CNN is of size 302: the 300-dimensional word embedding plus two bits for membership of the word in our style lexicons. Balancing parameter  $\lambda_c$  is set to 15. For sentiment task, we have used settings provided in (Shen et al., 2017).

**Implementation Details for Tag and Generate approach:** We use 4-layered transformers (Vaswani et al., 2017) to train both tagger and generator modules. Each transformer has 4 attention heads with a 512 dimensional embedding layer and hidden state size. Dropout (?) with p-value 0.3 is added for each layer in the transformer. For the politeness dataset the generator module is trained with data augmentation techniques like random word shuffle, word drops/replacements as proposed by (?). We empirically observed that these techniques provide an improvement in the fluency and diversity of the generations. Both modules were also trained with the BPE tokenization (?) using a vocabulary of size 16000 for all the datasets except for Captions, which was trained using 4000 BPE tokens. The value of the smoothing parameter  $\gamma$  in Eq. 3.11 is set to 0.75. For all datasets except Yelp we use phrases with  $p_1^2(w) \geq k = 0.9$  to construct  $\Gamma_2, \Gamma_1$  (§3.2.2). For Yelp  $k$  is set to 0.97. During inference we use beam search (beam size=5) to decode tagged sentences and targeted generations for tagger &

generator respectively. For the tagger, we re-rank the final beam search outputs based on the number of [TAG] tokens in the output sequence (favoring more [TAG] tokens).

## A.2 Additional Results

In Tables A.1, A.2, and A.3 we present the style transfer accuracy results broken-down to style categories. We denote the Cross-aligned Auto-Encoder model as CAE and our model as Back-translation for Style Transfer (BST).

Model	Style transfer	Acc	Style transfer	Acc
CAE	male → female	<b>64.75</b>	female → male	56.05
BST	male → female	54.59	female → male	<b>59.49</b>

TABLE A.1: Gender transfer accuracy.

Model	Style transfer	Acc	Style transfer	Acc
CAE	republican → democratic	65.44	democratic → republican	86.20
BST	republican → democratic	<b>80.55</b>	democratic → republican	<b>95.47</b>

TABLE A.2: Political slant transfer accuracy.

Model	Style transfer	Acc	Style transfer	Acc
CAE	negative → positive	81.63	positive → negative	79.65
BST	negative → positive	<b>95.68</b>	positive → negative	<b>81.65</b>

TABLE A.3: Sentiment modification accuracy.

In Table A.4, we detail the accuracy of the gender classifier on generated style-transferred sentences by an auto-encoder; Table A.5 shows the accuracy of transfer of political slant. We denote the Auto-Encoder as (AE) and our model as Back-translation for Style Transfer (BST).

Model	Style transfer	Acc	Style transfer	Acc
AE	male → female	41.48	female → male	41.88
BST	male → female	<b>54.59</b>	female → male	<b>59.49</b>

TABLE A.4: Gender transfer accuracy for Auto-encoder.

To evaluate the preservation of meaning by the Auto-Encoder, the experiments were setup as described in §3.3. We conducted four tests, each of 20 random samples for each type of style transfer. Note that we did not ask about appropriateness of the style transfer in this test, or



Model	Style transfer	Acc	Style transfer	Acc
AE	republican → democratic	60.76	democratic → republican	64.05
BST	republican → democratic	<b>80.55</b>	democratic → republican	<b>95.47</b>

TABLE A.5: Political slant transfer accuracy for Auto-encoder.

fluency of outputs, only about meaning preservation. We show the results of human evaluation in Table A.6

Style transfer	=	AE	BST
male → female	43.3	13.4	43.3
female → male	45.0	10.0	45.0
republican → democratic	43.3	3.4	<b>53.3</b>
democratic → republican	<b>55.00</b>	11.7	33.3

TABLE A.6: Human preference for meaning preservation in percentages.

### A.3 Examples of Generations

Examples of the original and style-transferred sentences generated by the baseline and our BST model are shown in Tables A.7, A.8 and A.9.

Input Sentence	CAE	BST
male → female		
my wife ordered country fried steak and eggs.	i got ta get the chicken breast.	my husband ordered the chicken salad and the fries.
great place to visit and maybe find that one rare item you just have never seen or can not find anywhere else.	we couldn't go back and i would be able to get me to get me.	great place to go back and try a lot of which you've never had to try or could not have been able to get some of the best.
the place is small but cosy and very clean	the staff and the place is very nice.	the place is great but very clean and very friendly.
female → male		
save yourself the huge headaches.	the sauces are excellent.	you are going to be disappointed.
would i discourage someone else from going?	i believe i would be back?	i wouldn't go back!
my husband ordered the salad and the dressing - lrb - blue cheese - rrb - was watered down.	the sauces - lrb - - rrb - - rrb - and - rrb -.	my wife ordered the macn- cheese and the salad - lrb - \$ 00 minutes - rrb - was cooked.

TABLE A.7: Gender style transfer examples. In addition to better preserving meaning, sentences generated by the BST model are generally grammatically better structured.

Input Sentence	CAE	BST
	republican → democratic	
i will continue praying for you and the decisions made by our government!	i am proud of you and your vote for us!	i will continue to fight for you and the rest of our democracy!
tom, i wish u would bring change.	i agree, senator warren and could be.	brian, i am proud to have you representing me.
all talk and no action-why dont you have some guts like breitbart	and then we will be praying for them and i am proud of this position and i am proud of	keep up and don't know, you have a lot of respect as breitbart
	democratic → republican	
as a hoosier, i thank you, rep. visclosky.	a lot , i am proud of you <unk>.	as a hoosier, i'm praying for you sir.
thank you for standing up for justice and against bigotry-racism, homophobia, sexism , misogyny, religious and xenophobia.	do you for standing up for highly and in bigotry-racism, programming, cut, granddaughters, unprecedented and excludes.	thanks for standing up for the constitution and get rid of obamacare, homophobie, cut, and actuality.
thank you for all you are doing for us, attorney general harris!	thank you for standing up for us and i am proud of us!	thanks lawmaker for all you do for us, senator scott!

TABLE A.8: Political slant style transfer examples. In addition to better preserving meaning, sentences generated by the BST model are generally grammatically better structured.

Input Sentence	CAE	BST
negative → positive		
crap fries, hard hamburger buns, burger tasted like crap!	good selection, fresh food, like like like!	empathy, the best food, but it was very nice!
the people behind the counter were not friendly whatsoever.	the people who the staff were friendly.	the people here are really good.
this place is bad news!	this place is great!	this place is amazing!
positive → negative		
the food is excellent and the service is exceptional!	the food is the food and the service is terrible.	the food is horrible and the service is terrible.
great as always, i love there food.	horrible as, i really don't eat here.	really disappointed, i couldn't be back.
i would recommend a visit here.	i would not recommend a dinner here.	i will not recommend this place.

TABLE A.9: Sentiment style transfer examples. In addition to better preserving meaning, sentences generated by the BST model are generally grammatically better structured.

## A.4 Preliminary experiments for Future Work

**Building Effective Style Representations:** We first experiment with simple averaging of BERT representations to get the style representation. We have dataset of sentences  $\mathbf{x} = \{x_1, \dots, x_n\}$  each  $x_i$  is mapped to one or more styles in the set  $\mathbf{y} = \{y_1, \dots, y_k\}$ . Suppose the set of sentences which belong to style  $y_i$  is  $\mathbf{x}_{y_i}$ . To build the representation of a style  $y_i \in \mathbf{y}$ , we follow:

$$\mathbf{S}_{y_i} = \Sigma_{x_j \in \mathbf{x}_{y_i}} \text{BERT}(x_j) \quad (\text{A.1})$$

We build representations for each style  $y_i \in \mathbf{y}$  using Eq. A.1. To test the quality of  $y_i$ , we design a binary classification task to determine if two sentences belong to the same style or not. Note that this is not a style classification task. We test our style representation by training three different classifiers for this task. All three models are based on the pre-trained base uncased BERT model (Devlin et al., 2019) and we don't fine tune the BERT layers. We get the representations of the two sentences ( $x_1$  and  $x_2$  say) using the BERT model ( $\mathbf{s}_1$  and  $\mathbf{s}_2$  say). For the *BERT-model* classifier, we concatenate  $[\mathbf{s}_1; \mathbf{s}_2]$  to get a representation  $h$  which then passed through two linear layers to get the final prediction. This model provides a baseline accuracy on how much can you learn about styles from BERT. For the *BERT-style* classifier, we subtract the style representations from the sentence representations. We get  $\mathbf{h}$  as follows:

$$\mathbf{h} = [\mathbf{s}_1 - \mathbf{S}_{y_i}; \mathbf{s}_2 - \mathbf{S}_{y_j}] \quad (\text{A.2})$$

where  $y_i$  and  $y_j$  are the styles of the two sentences  $x_1$  and  $x_2$  respectively. The *BERT-random* model obtains  $\mathbf{h}$  by subtracting the same random vector  $r$  from  $\mathbf{s}_1$  and  $\mathbf{s}_2$  and then concatenating them together. We have experimented with gender, age and education tasks from PASTEL dataset (Kang and Hovy, 2019) and the results are shown in Table A.10.

Model	Accuracy		
	Gender	Age	Education
<i>BERT-model</i>	56.10	58.42	58.94
<i>BERT-style</i>	54.49	56.61	50.96
<i>BERT-random</i>	54.97	58.10	58.94

TABLE A.10: Classifier accuracies

As we can see, after subtracting the style embeddings, we get a drop in classification accuracy suggesting that the style embeddings do capture some style information. This drop is not seen when we subtract a random vector which further provides evidence that the style embeddings capture information related to the style of the sentence.

One question that still remains is the usefulness of  $\mathbf{s}_1 - \mathbf{S}_{y_i}$  i.e the representation that remains after subtracting the style vector. We propose the following evaluation for assessing the quality of the style representation:

**Style Transfer:** We propose style transfer using the style representation obtained from Eq. A.1. This style representation could be concatenated to the input sentence representation to guide the generation process.

**Retrieval Techniques:** We design the following two retrieval experiments to test the style representation:

1. *Retrieve Style:* In this task, we take a sentence  $x_1$  and find  $k$  sentences with similar meaning to  $x_1$  in all the given styles using cosine similarity between their BERT representations. We average the BERT representations of these sentences and consider this as the meaning vector ( $m_1$ ) of  $x_1$ . We retrieve the style vector  $\mathbf{S}_y$  which is closest to  $\text{BERT}(x_1) - m_1$  using cosine similarity.
2. *Retrieve Sentence:* This task is performed to understand if the style representation can be used for style transformations. In this task, we take a the BERT representation of a sentence  $x_1$  (say  $s_1$ ). We get a transformed representation  $\hat{x}_1$  where  $\hat{x}_1 = s_1 - \mathbf{S}_{y_i} + \mathbf{S}_{y_j}$ , where  $y_i$  is the style of  $x_1$  and  $y_j$  is the style to which we would like to transform  $x_1$ . Our candidate set is made of  $k$  sentences from which  $k - 1$  belong to style  $y_i$  and one sentence belongs to  $y_j$ . The task is to retrieve the sentence that belongs to style  $y_j$  using cosine similarity between  $\hat{x}_1$  and the BERT representations of the candidate sentences. Note that the sentences selected for the candidate set will also be compared to  $x_1$  and only the sentences which are close in meaning with  $x_1$  would be chosen to belong to the candidate set.

**Structural Understanding.** Style may not necessarily be only defined by surface level features. We hypothesize that some styles are dependent heavily on the underlying structure of the sentence. In this case, we propose an ablation of the classifier performance with various structural features of the sentence. We plan to perform experiments for prediction of the style using only the POS tag sequences of the sentence. The POS tag N-grams denote the structure of the sentence. We have shown some preliminary results of classification accuracy in Table A.11 for the PASTEL dataset (Kang and Hovy, 2019) on the demographic attributes of gender, age and education. The results for the *BERT-model* are taken from (Kang and Hovy, 2019) and the *POS-model* corresponds to a SVM model trained only on POS unigram, bigram and trigram features. These results suggest that the POS N-grams are highly indicative of the gender styles of Male, Female and Non-binary. We also plan to perform similar experiments using features from parse trees, as well as additional features such as depth and breadth of the trees.

	<i>BERT-model</i>	<i>POS-model</i>
Gender	73.0	71.7
Age	46.3	40.5
Education	42.5	38.0

TABLE A.11: Comparison of classification accuracy for *BERT-model* and *POS-model*

## A.5 Details of Code

### Back-translation (BST) model

Github Link: <https://github.com/shrimai/Style-Transfer-Through-Back-Translation>

This link provide code base as well as trained models for the back-translation technique.

### Tag and Generate model

Github Link: <https://github.com/tag-and-generate/Politeness-Transfer-A-Tag-and-Generate-Approach>

This link provides the curated dataset for the politeness transfer task. It provides code as well as trained models for the tag and generate approach.

## Appendix B

# Appendix for Document Grounded Generation

This appendix details the hyper-parameters of the models described in Chapter 4 for all the tasks. It presents examples of the generated sentences for the Wikipedia edit generation task and examples of human dialogues collected for the grounded dialogue response generation task.

### B.1 Details of Training

#### Generative Models

##### Wikipedia Edit Generation

**Hyper-parameter settings:** For all our experiments with generative models, I have used bidirectional encoder, 2 layers in encoder and decoder, RNN size of 128, word vector size of 100. I have used sentencepiece toolkit<sup>1</sup> to use byte-pair-encoding (BPE) with a vocabulary size of 32k. I used stochastic gradient descent optimizer and the stopping criterion was perplexity on the validation set. I filtered our dataset to contain instances which have length of the document between 50 and 2000 tokens, length of the curated text between 20 and 500 tokens and the length of the update sentence between 5 and 200 tokens.

##### Dialogue Response Generation

**Experimental Setup:** For both COG and CRG model, I use a two-layer bidirectional LSTM as the encoder and a LSTM as the decoder. The dropout rate of the LSTM output is set to be

---

<sup>1</sup><https://github.com/google/sentencepiece>



0.3. The size of hidden units for both LSTMs is 300. I set the word embedding size to be 100, since the size of vocabulary is relatively small<sup>2</sup>. The models are trained with adam (Kingma and Ba, 2014) optimizer with learning rate 0.001 until they converge on the validation set for the perplexity criteria. I use beam search with size 5 for response generation. I use all the data (i.e all the conversations regardless of the rating and scenario) for training and testing. The proportion of train/validation/test split is 0.8/0.05/0.15.

### Pre-trained Encoder Decoder models

We use the transformer toolkit (Wolf et al., 2019) to implement the baseline and both CoDR and DoHA models.<sup>3</sup> Both DoHA (§4.2.3) and CoDR (§4.2.3) have the same dimensions and architecture of the BART model (?). For the DoHA model, we initialize *CrossAttention\_Doc* with same pre-trained weights of *CrossAttention*. Hence, the layer size of the *CrossAttention\_Doc* multi-head is the same as the layer size of *CrossAttention* multi-head in BART. We experimented with two learning rates 5e-5 and 2e-5. We report numbers for the best trained models in each case. Specifically, we report numbers with 5e-5 learning rate for DoHA and CoDR models on the CMU\_DoG dataset and the BART baseline for all the three datasets. For Wikipedia Update Generation and Wizard of Wikipedia dataset, we choose the DoHA and CoDR models trained with 2e-5 learning rate. We maintain a common environment (in terms of GPU, operating system, Pytorch version and transformer version) to run all the experiments. We train all the models for 25 epochs.

Zhao et al. (2020a) numbers are directly taken from the paper as the pre-trained model or the generated outputs are not available. We use the same data splits and evaluation toolkits for comparable setting. Hence, Rouge-L and Meteor values are not available for this model. The BLEU, Meteor and Rouge-L numbers are different from (Prabhumoye et al., 2019b) due to the usage of different tool-kits in measuring their values.

**Convergence:** Figures B.1 and B.2 shows the convergence of the baseline BART model in comparison with the CoDR and DoHA models on the development sets of CMU\_DoG and Wizard of Wikipedia respectively. We observe that at same number of updates, DoHA and CoDR perform better than BART. This is especially relevant for big datasets like Wikipedia Update Generation which take 15 days to complete 25 epochs.

<sup>2</sup>The total number of tokens is 46000, and we limit the vocabulary to be 10000 tokens.

<sup>3</sup>The results are subject to changes in the codebase of the toolkit. Note that we will release our code and trained models to ensure reproducibility of results.

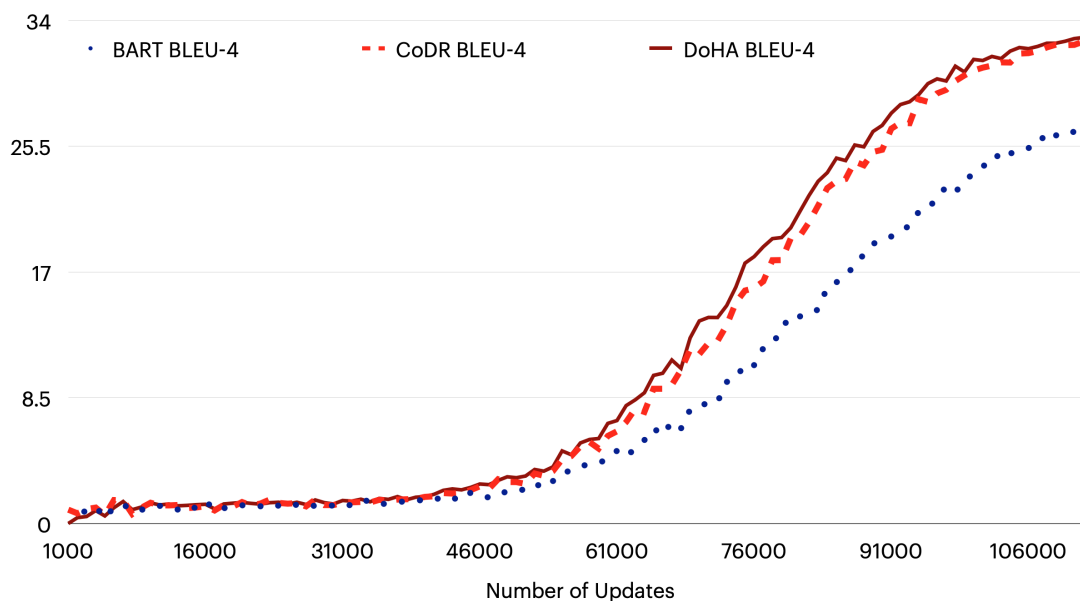


FIGURE B.1: Convergence of CMU\_DoG development data on the automated metric.

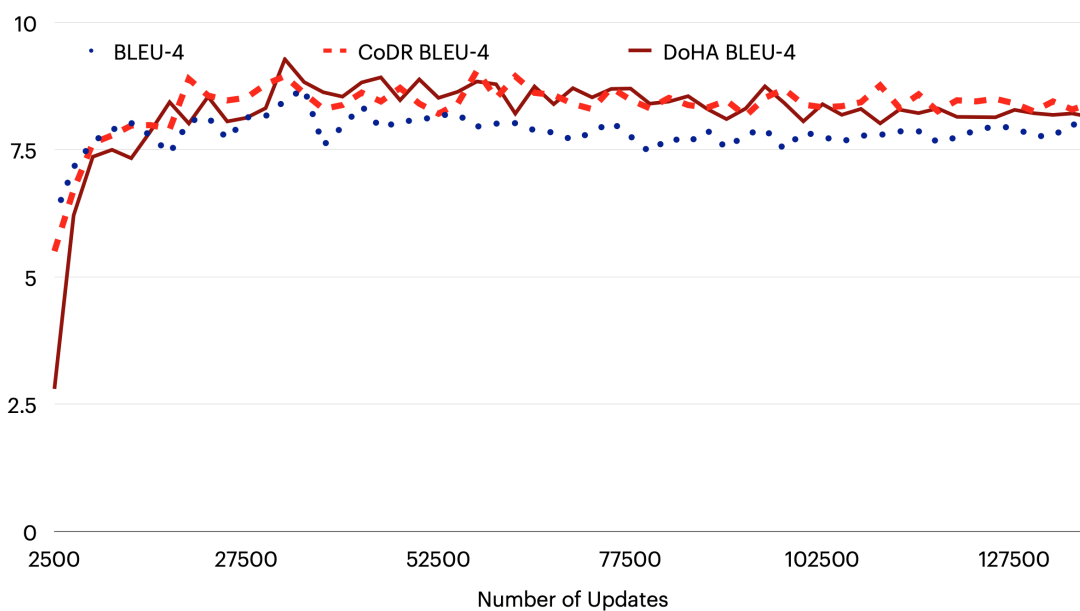


FIGURE B.2: Convergence of CMU\_DoG development data on the automated metric.

## B.2 Additional Dataset Details

Table B.1 shows the maximum sequence lengths used for all the three datasets for both source and target. The data statistics are shown in Table B.2.<sup>4</sup>

<sup>4</sup>We try to closely follow the processing of the original papers for each of the three datasets.

<b>Dataset</b>	<b>Source Len</b>	<b>Target Len</b>
Wikipedia Update Generation	1024	128
CMU_DoG dataset	512	128
Wizard of Wikipedia	900	40

TABLE B.1: Sequence Lengths

<b>Dataset</b>	<b>Train</b>	<b>Dev</b>	<b>Test</b>
Wikipedia Update Generation	580.0k	6.0k	50.0k
CMU_DoG	72.9k	4.8k	13.2k
Wizard of Wikipedia	166.7k	17.7k	8.7k

TABLE B.2: Dataset Statistics

### Additional Details for CMU\_DoG

**Movie lists:** Here is a list of movie that were selected as topics of conversation.

- Batman Begins
- Bruce Almighty
- Batman v Superman: Dawn of Justice
- Catch me if you can
- Despicable me (2010)
- Dunkirk
- Frozen (2013)
- Home Alone
- How to Train Your Dragon (2010)
- The Imitation Game
- Iron Man (2008)
- Jaws
- John Wick (2014)
- La La Land
- Maleficent
- Mean Girls
- Monsters University
- Real Steel
- The Avengers (2012)
- The Blind Side
- The Great Gatsby (2013)
- The Inception
- The Notebook
- The Post
- The Shape of Water
- The Social Network
- The Wolf of Wall Street
- Toy Story
- Wonder Woman
- Zootopia

## Instructions given to the workers

### Scenario 1: users with document

- The user you are pairing does not have the document you hold. Please read the document first.
- Tell the other user what the movie is, and try to persuade the other user to watch/not to watch the movie using the information in the document.
- You should try to discuss the new paragraph when the document has changed.
- You will have 3 turns of conversation with your partner on each of the documents.
- You will be given 4 documents each containing a short paragraph. The new paragraph might show just beneath the previous document.
- The next document will be loaded automatically after you finish 3 turns discussing the current document.
- You cannot use information you personally know that is not included there. You can use any information given in the document in the conversation.

### Scenario 1: users without document

- The other user will read a document about a movie.
- If you are not told the name of the movie, try to ask the movie name.
- After you are told the name of the movie, pretend you are interested in watching the movie, and try to gather all the information you need to make a decision whether to watch the movie in the end.
- You don't have to tell the other user your decision in the end, but please share your mind at the feedback page.

### Scenario 2: both users with document

- The user you pair with has the same set of documents as yours. Please read the document first
- Imagine you just watched this movie. Discuss the content in the document with the other user, and show whether you like/dislike the movie.
- You should try to discuss the new paragraph when the document has changed.
- You will have 3 turns of conversation with your partner on each of the documents.

- You will be given 4 documents each containing a short paragraph. The new paragraph might show just beneath the previous document
- The next document will be loaded automatically after you finish 3 turns discussing the current document.
- You cannot use information you personally know that is not included there. You can use any information given in the document in the conversation.

## **Post conversation survey questions**

### **For users with document**

Choose any:

- The document is understandable.
- The other user is actively responding to me.
- The conversation goes smoothly.

Choose one of the following:

- I have watched the movie before.
- I have not watched the movie before.

### **For users without document**

Choose any:

- The document is understandable.
- The other user is actively responding to me.
- The conversation goes smoothly.

Choose one of the following:

- I will watch the movie after the other user's introduction.
- I will not watch the movie after the other user's introduction.

## Conversation Example 1

This is an example of conversation which follows *Scenario 1* where *user2* has access to sections. Tables [B.3](#), [B.4](#), [B.5](#) and [B.6](#) shows the conversation corresponding to each of the four sections of the document.

## Conversation Example 2

This is an example of conversation which follows *Scenario 2* where both users have access to sections. Tables [B.7](#), [B.8](#), [B.9](#) and [B.10](#) shows the conversation corresponding to each of the four sections of the document.

Section 1	
<b>Name</b>	The inception
<b>Year</b>	2009
<b>Director</b>	Christopher Nolan
<b>Genre</b>	scientific
<b>Cast</b>	Leonardo DiCaprio as Dom Cobb, a professional thief who specializes in conning secrets from his victims by infiltrating their dreams. Joseph Gordon-Levitt as Arthur, Cobb's partner who manages and researches the missions. Ellen Page as Ariadne, a graduate student of architecture who is recruited to construct the various dreamscapes, which are described as mazes. Tom Hardy as Eames, a sharp-tongued associate of Cobb.
<b>Critical Response</b>	wildly ingenious chess game, the result is a knockout. DiCaprio, who has never been better as the tortured hero, draws you in with a love story that will appeal even to non-sci-fi fans. I found myself wishing Inception were weirder, further out the film is Nolan's labyrinth all the way, and it's gratifying to experience a summer movie with large visual ambitions and with nothing more or less on its mind than (as Shakespeare said) a dream that hath no bottom. Have no idea what so many people are raving about. It's as if someone went into their heads while they were sleeping and planted the idea that Inception is a visionary masterpiece and hold on Whoa! I think I get it. The movie is a metaphor for the power of delusional hype a metaphor for itself.
<b>Introduction</b>	Dominick Cobb and Arthur are extractors, who perform corporate espionage using an experimental military technology to infiltrate the subconscious of their targets and extract valuable information through a shared dream world. Their latest target, Japanese businessman Saito, reveals that he arranged their mission himself to test Cobb for a seemingly impossible job: planting an idea in a person's subconscious, or inception. To break up the energy conglomerate of ailing competitor Maurice Fischer, Saito wants Cobb to convince Fischer's son and heir, Robert, to dissolve his father's company.
<b>Rating</b>	Rotten Tomatoes: 86% and average: 8.1/10; IMDB: 8.8/10
Conversation	
user2:	Hey have you seen the inception?
user1:	No, I have not but have heard of it. What is it about
user2:	It's about extractors that perform experiments using military technology on people to retrieve info about their targets.
user1:	Sounds interesting do you know which actors are in it?
user2:	I haven't watched it either or seen a preview. Bu5 it's scifi so it might be good. Ugh Leonardo DiCaprio is the main character
user2:	He plays as Don Cobb
user1:	Oh okay, yeah I'm not a big scifi fan but there are a few movies I still enjoy in that genre.
user1:	Is it a long movie?
user2:	Doesn't say how long it is.
user2:	The Rotten Tomatoes score is 86%

TABLE B.3: Utterances that corresponds to section 1 of the document in the example conversation 1.

Section 2	
<b>Scene 1</b>	When the elder Fischer dies in Sydney, Robert Fischer accompanies the body on a ten-hour flight back to Los Angeles, which the team (including Saito, who wants to verify their success) uses as an opportunity to sedate and take Fischer into a shared dream. At each dream level, the person generating the dream stays behind to set up a 'kick' that will be used to awaken the other sleeping team members from the deeper dream level; to be successful, these kicks must occur simultaneously at each dream level, a fact complicated due to the nature of time which flows much faster in each successive level. The first level is Yusuf's dream of a rainy Los Angeles. The team abducts Fischer, but they are attacked by armed projections from Fischer's subconscious, which has been specifically trained to defend him against such intruders. The team takes Fischer and a wounded Saito to a warehouse, where Cobb reveals that while dying in the dream would normally wake Saito up, the powerful sedatives needed to stabilize the multi-level dream will instead send a dying dreamer into 'limbo', a world of infinite subconscious from which escape is extremely difficult, if not almost impossible, and a dreamer risks forgetting they are in a dream. Despite these setbacks, the team continues with the mission.
Conversation	
user1:	Wow, that's impressive. I like to look at Rotten Tomatoes when debating whether or not to see a movie. Do you know the director?
user2:	Something about Dom Cobb infiltrates peoples dreams in a dream world.
user2:	The director is Christopher nolan
user2:	Heard of him?
user2:	Wow I thought this was recent but it came out in 2009.
user1:	He directed The Dark Knight which I enjoy. Yeah, I know it's been out awhile but 2009 does seem to be a while back now. Time flies.
user1:	Do you know if it won any awards?
user1:	or how much it made at the box office?
user2:	Oh wow I loved the dark night movies. And it doesn't say if it's won awards or how much at box office.
user2:	A critic did say it could be "weirder"

TABLE B.4: Utterances that corresponds to section 2 of the document in the example conversation 1.



---

Section 3	
<b>Scene 2</b>	Cobb reveals to Ariadne that he and Mal went to Limbo while experimenting with the dream-sharing technology. Sedated for a few hours of real time, they spent fifty years in a dream constructing a world from their shared memories. When Mal refused to return to reality, Cobb used a rudimentary form of inception by reactivating her totem (an object dreamers use to distinguish dreams from reality) and reminding her subconscious that their world was not real. However, when she woke up, Mal still believed that she was dreaming. In an attempt to 'wake up' for real, Mal committed suicide and framed Cobb for her death to force him to do the same. Facing a murder charge, Cobb fled the U.S., leaving his children in the care of Professor Miles.
Conversation	
user1:	The concept seems interesting and it has a good lead actor as well as director and reviews. I think it must be good. The plot does seem weird, that's for sure.
user2:	Tom Hardy is in the movie as the character Earnes. And yeah the plot is a bit strange.
user2:	I might watch this movie now.
user1:	I think I may as well. I can't say I've heard of Tom Hardy however. Is there any other supporting actors?
user2:	Oh Earnes is a sharp tongue associate of Cobb.
user2:	Ellen Page
user1:	Oh, cool. I am familiar with her. She's in a number of good movies and is great.
user2:	She plays Ariadne , she is a graduate student that constructs the dreamscapes, they're like mazes.

---

TABLE B.5: Utterances that corresponds to section 3 of the document in the example conversation 1.

Section 4	
<b>Scene 3</b>	Through his confession, Cobb makes peace with his guilt over Mal's death. Ariadne kills Mal's projection and wakes Fischer up with a kick. Revived at the mountain hospital, Fischer enters a safe room to discover and accept the planted idea: a projection of his dying father telling him to be his own man. While Cobb remains in Limbo to search for Saito, the other team members ride the synchronized kicks back to reality. Cobb eventually finds an aged Saito in Limbo and reminds him of their agreement. The dreamers all awake on the plane and Saito makes a phone call. Upon arrival at Los Angeles Airport, Cobb passes the U.S. immigration checkpoint and Professor Miles accompanies him to his home. Using his totem a spinning top that spins indefinitely in a dream world but falls over in reality Cobb conducts a test to prove that he is indeed in the real world, but he ignores its result and instead joins his children in the garden.
Conversation	
user1:	Hmm interesting. Do you know if it's an action movie or mostly just scifi?
user2:	Says scientific
user1:	Certainly seems unique. Do you know if it is based off a book or a previous work?
user2:	Something about at the end he has trouble determining which is reality and which is a dream. It doesn't say it's based off anything.
user1:	Sounds like it might be suspense/thriller as well as scifi which is cool. It seems pretty confusing but enticing. Makes me want to see it to try and figure it all out.
user2:	Yeah its like its got a bit of mystery too. Trying to figure out what's real and what's not.
user1:	I can't think of any other movie or even book that has a related story either which makes it very interesting. A very original concept.
user2:	Yeah well have great day. :)

TABLE B.6: Utterances that corresponds to section 4 of the document in the example conversation 1.

Section 1	
<b>Name</b>	The Shape of Water
<b>Year</b>	2017
<b>Director</b>	Guillermo del Toro
<b>Genre</b>	Fantasy, Drama
<b>Cast</b>	Sally Hawkins as Elisa Esposito, a mute cleaner who works at a secret government laboratory. Michael Shannon as Colonel Richard Strickland, a corrupt military official, Richard Jenkins as Giles, Elisa's closeted neighbor and close friend who is a struggling advertising illustrator. Octavia Spencer as Zelda Delilah Fuller, Elisa's co-worker and friend who serves as her interpreter. Michael Stuhlbarg as Dimitri Mosenkov, a Soviet spy working as a scientist studying the creature, under the alias Dr. Robert Hoffstetler.
<b>Critical Response</b>	one of del Toro's most stunningly successful works, also a powerful vision of a creative master feeling totally, joyously free. Even as the film plunges into torment and tragedy, the core relationship between these two unlikely lovers holds us in thrall. Del Toro is a world-class film artist. There's no sense trying to analyze how he does it. The Shape of Water has tenderness uncommon to del Toro films. While The Shape of Water isn't groundbreaking, it is elegant and mesmerizing. refer Sally Hawkins' mute character as 'mentally handicapped' and for erroneously crediting actor Benicio del Toro as director.
<b>Introduction</b>	The Shape of Water is a 2017 American fantasy drama film directed by Guillermo del Toro and written by del Toro and Vanessa Taylor. It stars Sally Hawkins, Michael Shannon, Richard Jenkins, Doug Jones, Michael Stuhlbarg, and Octavia Spencer. Set in Baltimore in 1962, the story follows a mute custodian at a high-security government laboratory who falls in love with a captured humanoid amphibian creature.
<b>Rating</b>	Rotten Tomatoes: 92% and average: 8.4/10 Metacritic Score: 87/100 CinemaScore: A
Conversation	
user1:	Hi
user2:	Hi
user2:	I thought The Shape of Water was one of Del Toro's best works. What about you?
user1:	Did you like the movie?
user1:	Yes, his style really extended the story.
user2:	I agree. He has a way with fantasy elements that really helped this story be truly beautiful.
user2:	It has a very high rating on rotten tomatoes, too. I don't always expect that with movies in this genre.
user1:	ally Hawkins acting was phenomenally expressive. Didn't feel her character was mentally handicapped.
user2:	The characterization of her as such was definitely off the mark.

TABLE B.7: Utterances that corresponds to section 1 of the document in the example conversation 2.

---

Section 2	
<b>Scene 1</b>	Elisa Esposito, who as an orphaned child, was found in a river with wounds on her neck, is mute, and communicates through sign language. She lives alone in an apartment above a cinema, and works as a cleaning-woman at a secret government laboratory in Baltimore at the height of the Cold War. Her friends are her closeted next-door neighbor Giles, a struggling advertising illustrator who shares a strong bond with her, and her co-worker Zelda, a woman who also serves as her interpreter at work. The facility receives a mysterious creature captured from a South American river by Colonel Richard Strickland, who is in charge of the project to study it. Curious about the creature, Elisa discovers it is a humanoid amphibian. She begins visiting him in secret, and the two form a close bond.
Conversation	
user1:	Might as well label Giles too.
user2:	haha. because he is closeted?
user2:	Whoever made that comment was certainly not well informed and not politically correct by any stretch.
user1:	I think Octavia Spencer should look for more roles set in the early 60s.
user2:	Do you think that the creature they find in the movie is supposed to be somehow connected to the cold war?

---

TABLE B.8: Utterances that corresponds to section 2 of the document in the example conversation 2.

---

Section 3	
<b>Scene 2</b>	Elisa keeps the creature in her bathtub, adding salt to the water to keep him alive. She plans to release the creature into a nearby canal when it will be opened to the ocean in several days' time. As part of his efforts to recover the creature, Strickland interrogates Elisa and Zelda, but the failure of his advances toward Elisa hampers his judgment, and he dismisses them. Back at the apartment, Giles discovers the creature devouring one of his cats, Pandora. Startled, the creature slashes Giles's arm and rushes out of the apartment. The creature gets as far as the cinema downstairs before Elisa finds him and returns him to her apartment. The creature touches Giles on his balding head and his wounded arm; the next morning, Giles discovers his hair has begun growing back and the wounds on his arm have healed. Elisa and the creature soon become romantically involved, having sex in her bathroom, which she at one point fills completely with water.
Conversation	
user1:	Actually Del Toro does an incredible job showing working people.
user2:	That's an excellent point.
user1:	Yes, the Cold War invented the Russians, I kind of thought it also represented technology in general.
user2:	That makes perfect sense.
user2:	I really like that Eliza chose to keep the creature in her bathtub.
user1:	It was interesting that neither power treated the monster well.
user1:	Yes the magical realism was truly magical ... easy to suspend disbelief.

---

TABLE B.9: Utterances that corresponds to section 3 of the document in the example conversation 2.

Section 4	
<b>Scene 3</b>	Hoyt gives Strickland an ultimatum, asking him to recover the creature within 36 hours. Meanwhile, Mosenkov is told by his handlers that he will be extracted in two days. As the planned release date approaches, the creature's health starts deteriorating. Mosenkov leaves to rendezvous with his handlers, with Strickland tailing him. At the rendezvous, Mosenkov is shot by one of his handlers, but Strickland shoots the handlers dead and then tortures Mosenkov for information. Mosenkov implicates Elisa and Zelda before dying from his wounds. Strickland then threatens Zelda in her home, causing her terrified husband to reveal that Elisa had been keeping the creature. Strickland searches Elisa's apartment and finds a calendar note revealing when and where she plans to release him. At the canal, Elisa and Giles bid farewell to the creature, but Strickland arrives and attacks them all. Strickland knocks Giles down and shoots the creature and Elisa, who both appear to die. However, the creature heals himself and slashes Strickland's throat, killing him. As police arrive on the scene with Zelda, the creature takes Elisa and jumps into the canal, where, deep under water, he heals her. When he applies his healing touch to the scars on her neck, she starts to breathe through gills. In a closing voiceover narration, Giles conveys his belief that Elisa lived 'happily ever after' with the creature.
Conversation	
user2:	Yes. I think it was beautiful that the creature essentially had healing power.
user1:	Del Toro does well with violence.
user1:	The ending was suspenseful, without being over the top.
user2:	What a powerful ending. Even though it was obviously a pure fantasy scenario, there was so much real emotion.
user2:	He does do well with violence. I've noticed that in all of his movies.
user2:	Del Toro is one of my favorite directors.
user1:	Yes, happy endings usually feel fake. This one felt great.
user2:	Totally. It felt like what should have happened, rather than just a sappy pretend ending that was forced on the viewer.
user1:	Mine too. Evidently Hollywood is starting to agree.
user2:	It took a while, but yes, finally.
user1:	It really appeared to be filmed in Baltimore. Installation looked so authentic.
user2:	Do you know where it was actually filmed?
user1:	No. Can you imagine soaking in that pool?
user2:	:)
user1:	Would make a great tourist draw.
user2:	That would be amazing! What a great idea!
user2:	Haven't we completed the amount of discussion needed yet?
user1:	Place looked like a cross between a nuclear power plant and an aquarium. I think we hit all the points mentioned.

TABLE B.10: Utterances that corresponds to section 4 of the document in the example conversation 2.

## B.3 Details of Code

### Wikipedia Update Generation

Github Link: <https://github.com/shrimai/Towards-Content-Transfer-through-Grounded-Text-Generation>

This link contains the dataset available for download as well as code and pre-trained models for the generative models described in §4.2.1.

### CMU\_DoG Dataset

Github Link: [https://github.com/festvox/datasets-CMU\\_DoG](https://github.com/festvox/datasets-CMU_DoG)

This link contains the raw conversations data as well as processed train, dev and test splits for the task. It also contains the list of movies and the documents used in collecting the dataset.

### Pre-trained Encoder Decoder models

Github Link: <https://github.com/shrimai/Focused-Attention-Improves-Documents-Grounded-Generation>

This link contains the code base for the models described in §4.2.3 which include the BART, CoDR and DoHA model. It also includes models trained on the three datasets described in chapter 4.

## Appendix C

# Appendix for Sentence Ordering

This appendix details the hyper-parameters of the models described in Chapter 5 and presents examples of the orders predicted for SIND and NIPS datasets by the B-TSort and the B-AON models.

### C.1 Details of Training

**Hyper-parameters.** For AON model we use the code base provided by the authors in (Cui et al., 2018) and we maintain the hyper-parameters described in the paper. For the paragraph encoder of the B-AON models, we follow the same scheme of the AON model but for its sentence encoder we use hyper-parameters of the BERT setting. We use the pretrained BERT uncased base model with 12 layers for the B-AON and B-TSORT models. We fine-tune the BERT model in both cases. Hence, we replace the Adadelta optimizer with the BertAdam (Wolf et al., 2019) optimizer for the B-AON model. The LSTMs in the L-TSort model uses an RNN size of 512 and it uses the same vocabularies as the AON model. L-TSort is trained using stochastic gradient descent with dropout of 0.2, learning rate of 1.0 and learning decay rate of 0.5. For B-TSort and L-TSort we use accuracy on the validation set to stop training. For B-TSort and B-AON we use learning rate of  $5e-5$  with adam epsilon value of  $1e-8$ .

### C.2 Examples of Sentence Order Predictions

Gold Order	B-TSort Order	B-AON Order
<p>the family sits together for dinner on the first night of the annual reunion. the restaurant we chose had amazing food and everyone loved the presentation. gemma really adored the restaurants decorations and was always gazing at them. aunt harriot had a little trouble deciding what kind of wine she wanted tonight. bob had the whole family cracking up with his jokes.</p>	<p>the family sits together for dinner on the first night of the annual reunion. the restaurant we chose had amazing food and everyone loved the presentation. aunt harriot had a little trouble deciding what kind of wine she wanted tonight. gemma really adored the restaurants decorations and was always gazing at them. bob had the whole family cracking up with his jokes.</p>	<p>the family sits together for dinner on the first night of the annual reunion. aunt harriot had a little trouble deciding what kind of wine she wanted tonight. bob had the whole family cracking up with his jokes. gemma really adored the restaurants decorations and was always gazing at them. the restaurant we chose had amazing food and everyone loved the presentation.</p>
<p>he wanted to take a ride on his new bike. we went on a nice ride out to the lake. we really enjoyed the beautiful view from the dock. it was very peaceful watching the boats. we had such a busy day he needed a nap.</p>	<p>we went on a nice ride out to the lake. he wanted to take a ride on his new bike. we really enjoyed the beautiful view from the dock. it was very peaceful watching the boats. we had such a busy day he needed a nap.</p>	<p>we went on a nice ride out to the lake. he wanted to take a ride on his new bike. it was very peaceful watching the boats. we really enjoyed the beautiful view from the dock. we had such a busy day he needed a nap.</p>
<p>when we finally brought our son home from the hospital so many people were at home with us to see him. everyone wanted a chance to hold him! we were all so happy to have a new addition to the family. my parents were so proud to be grand parents! i am so happy and i love my son very much!</p>	<p>when we finally brought our son home from the hospital so many people were at home with us to see him. we were all so happy to have a new addition to the family. everyone wanted a chance to hold him! my parents were so proud to be grand parents! i am so happy and i love my son very much!</p>	<p>my parents were so proud to be grand parents! when we finally brought our son home from the hospital so many people were at home with us to see him. we were all so happy to have a new addition to the family. everyone wanted a chance to hold him! i am so happy and i love my son very much!</p>

TABLE C.1: Examples of predicted sentence orders for B-TSort and B-AON model for SIND dataset.



Gold Order	B-TSort Order	B-AON Order
<p>we study how well one can recover sparse principal components of a data matrix using a sketch formed from a few of its elements. we show that for a wide class of optimization problems, if the sketch is close (in the spectral norm) to the original data matrix, then one can recover a near optimal solution to the optimization problem by using the sketch. in particular, we use this approach to obtain sparse principal components and show that for <math>m</math> data points in <math>n</math> dimensions, <math>\mathcal{O}(-2k \max(m, n))</math> elements gives an <math>\epsilon</math>-additive approximation to the sparse pca problem (<math>k</math> is the stable rank of the data matrix). we demonstrate our algorithms extensively on image, text, biological and financial data. the results show that not only are we able to recover the sparse pcas from the incomplete data, but by using our sparse sketch, the running time drops by a factor of five or more.</p>	<p>we study how well one can recover sparse principal components of a data matrix using a sketch formed from a few of its elements. we show that for a wide class of optimization problems, if the sketch is close (in the spectral norm) to the original data matrix, then one can recover a near optimal solution to the optimization problem by using the sketch. in particular, we use this approach to obtain sparse principal components and show that for <math>m</math> data points in <math>n</math> dimensions, <math>\mathcal{O}(-2k \max(m, n))</math> elements gives an <math>\epsilon</math>-additive approximation to the sparse pca problem (<math>k</math> is the stable rank of the data matrix). the results show that not only are we able to recover the sparse pcas from the incomplete data, but by using our sparse sketch, the running time drops by a factor of five or more. we demonstrate our algorithms extensively on image, text, biological and financial data.</p>	<p>we study how well one can recover sparse principal components of a data matrix using a sketch formed from a few of its elements. in particular, we use this approach to obtain sparse principal components and show that for <math>m</math> data points in <math>n</math> dimensions, <math>\mathcal{O}(-2k \max(m, n))</math> elements gives an <math>\epsilon</math>-additive approximation to the sparse pca problem (<math>k</math> is the stable rank of the data matrix). we show that for a wide class of optimization problems, if the sketch is close (in the spectral norm) to the original data matrix, then one can recover a near optimal solution to the optimization problem by using the sketch. the results show that not only are we able to recover the sparse pcas from the incomplete data, but by using our sparse sketch, the running time drops by a factor of five or more. we demonstrate our algorithms extensively on image, text, biological and financial data.</p>
<p>we develop a latent variable model and an efficient spectral algorithm motivated by the recent emergence of very large data sets of chromatin marks from multiple human cell types . a natural model for chromatin data in one cell type is a hidden markov model ( hmm ) ; we model the relationship between multiple cell types by connecting their hidden states by a fixed tree of known structure . the main challenge with learning parameters of such models is that iterative methods such as em are very slow , while naive spectral methods result in time and space complexity exponential in the number of cell types . we exploit properties of the tree structure of the hidden states to provide spectral algorithms that are more computationally efficient for current biological datasets . we provide sample complexity bounds for our algorithm and evaluate it experimentally on biological data from nine human cell types . finally , we show that beyond our specific model , some of our algorithmic ideas can be applied to other graphical models .</p>	<p>a natural model for chromatin data in one cell type is a hidden markov model ( hmm ) ; we model the relationship between multiple cell types by connecting their hidden states by a fixed tree of known structure . the main challenge with learning parameters of such models is that iterative methods such as em are very slow , while naive spectral methods result in time and space complexity exponential in the number of cell types . we develop a latent variable model and an efficient spectral algorithm motivated by the recent emergence of very large data sets of chromatin marks from multiple human cell types . we exploit properties of the tree structure of the hidden states to provide spectral algorithms that are more computationally efficient for current biological datasets . we provide sample complexity bounds for our algorithm and evaluate it experimentally on biological data from nine human cell types . finally , we show that beyond our specific model , some of our algorithmic ideas can be applied to other graphical models .</p>	<p>the main challenge with learning parameters of such models is that iterative methods such as em are very slow , while naive spectral methods result in time and space complexity exponential in the number of cell types . a natural model for chromatin data in one cell type is a hidden markov model ( hmm ) ; we model the relationship between multiple cell types by connecting their hidden states by a fixed tree of known structure . we develop a latent variable model and an efficient spectral algorithm motivated by the recent emergence of very large data sets of chromatin marks from multiple human cell types . we exploit properties of the tree structure of the hidden states to provide spectral algorithms that are more computationally efficient for current biological datasets . we provide sample complexity bounds for our algorithm and evaluate it experimentally on biological data from nine human cell types . finally , we show that beyond our specific model , some of our algorithmic ideas can be applied to other graphical models .</p>

TABLE C.2: Examples of predicted sentence orders for B-TSort and B-AON model for NIPS dataset.

### C.3 Details of Code

Github Link: <https://github.com/shrimai/Topological-Sort-for-Sentence-Ordering>

This link provides the code base and trained models for the B-TSort and L-TSort methods described in §5.1.

# Bibliography

- Nikolaos Aletras, Elliott Ash, Leslie Barrett, Daniel Chen, Adam Meyers, Daniel Preotiuc-Pietro, David Rosenberg, and Amanda Stent, editors. 2019. *Proceedings of the Natural Legal Language Processing Workshop 2019*. Association for Computational Linguistics, Minneapolis, Minnesota.
- Larry Alexander and Michael Moore. 2016. Deontological Ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, winter 2016 edition. Metaphysics Research Lab, Stanford University.
- Haifa Alharthi and Diana Inkpen. 2019. Study of linguistic features incorporated in a literary book recommender system. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1027–1034.
- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. [Publicly available clinical BERT embeddings](#). In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Gabor Angeli and Christopher Manning. 2013. [Philosophers are mortal: Inferring the truth of unseen facts](#). In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 133–142, Sofia, Bulgaria. Association for Computational Linguistics.
- Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354.
- Aristotle. 350 B.C.E. *Nicomachean Ethics*.
- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018*.
- Mariam Asad, Lynn Dombrowski, Sasha Costanza-Chock, Sheena Erete, and Christina Harrington. 2019. Academic accomplices: Practical strategies for research justice. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*, pages 353–356.

- David Atkinson, Kumar Bhargav Srinivasan, and Chenhao Tan. 2019. What gets echoed? understanding the “pointers” in explanations of persuasive arguments. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2904–2914.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.
- Mona Baker. 1992. *In Other Words: A Coursebook on Translation*. Routledge, United Kingdom.
- Vidhisha Balachandran, Artidoro Pagnoni, Jay Yoon Lee, Dheeraj Rajagopal, Jaime Carbonell, and Yulia Tsvetkov. 2020. [Structsum: Incorporating latent and explicit sentence dependencies for single document summarization](#). *ArXiv e-prints*.
- David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.
- Eva Banik, Claire Gardent, and Eric Kow. 2013. [The KBGen challenge](#). In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 94–97, Sofia, Bulgaria. Association for Computational Linguistics.
- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. [Findings of the 2019 conference on machine translation \(WMT19\)](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.
- Regina Barzilay and Noemie Elhadad. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*.
- Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.
- Regina Barzilay and Lillian Lee. 2004. [Catching the drift: Probabilistic content models, with applications to generation and summarization](#). In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 113–120, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Regina Barzilay and Kathleen R. McKeown. 2005. [Sentence fusion for multidocument news summarization](#). *Computational Linguistics*, 31(3):297–328.
- Christine Basta, Marta R Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39.

- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.
- Emily M. Bender, Dirk Hovy, and Alexandra Schofield. 2020. [Integrating ethics into the NLP curriculum](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–9, Online. Association for Computational Linguistics.
- Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons.
- Christina L Bennett. 2005. Large scale evaluation of corpus-based synthesizers: Results and lessons from the blizzard challenge 2005. In *Ninth European Conference on Speech Communication and Technology*.
- Jeremy Bentham. 1789. *An introduction to the principles of morals and legislation*. Clarendon Press.
- Jeremy Bentham. 1843. *The Rationale of Reward*.
- Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. [Ethical research protocols for social media health research](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.
- Shane Bergsma and Benjamin Van Durme. 2013. Using conceptual class attributes to characterize social media users. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 710–720.
- Parminder Bhatia, Steven Lin, Rashmi Gangadharaiah, Byron Wallace, Izhak Shafran, Chaitanya Shivade, Nan Du, and Mona Diab, editors. 2020. *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*. Association for Computational Linguistics, Online.
- Robert J Bies. 2001. Interactional (in) justice: The sacred and the profane. *Advances in organizational justice*, 89118.
- Robert J Bies. 2015. Interactional justice: Looking backward, looking forward. *The Oxford Handbook of Justice in the Workplace*, page 89.
- Alan W Black. 2018. Computational propaganda. [http://demo.clab.cs.cmu.edu/ethical\\_nlp/slides/12\\_Propaganda.pdf](http://demo.clab.cs.cmu.edu/ethical_nlp/slides/12_Propaganda.pdf) (accessed April 9, 2021).
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. [Language \(technology\) is power: A critical survey of “bias” in NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

- Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proc. EMNLP*.
- L.P. Bloomberg. 1981. Bloomberg professional services. [https://www.bloomberg.com/professional/solution/content-and-data/?utm\\_source=bloomberg-menu&utm\\_medium=bcom&bbgsum=DG-WS-PROF-SOLU-DATACONT-bbgmenu](https://www.bloomberg.com/professional/solution/content-and-data/?utm_source=bloomberg-menu&utm_medium=bcom&bbgsum=DG-WS-PROF-SOLU-DATACONT-bbgmenu) (accessed April 9, 2021).
- Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.
- Antoine Bosselut and Yejin Choi. 2019. Dynamic knowledge graph construction for zero-shot commonsense question answering. *arXiv preprint arXiv:1911.03876*.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019a. **COMET: Commonsense transformers for automatic knowledge graph construction**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019b. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.
- Abdelghani Bouziane, D. Bouchiha, Noureddine Doumi, and M. Malki. 2015. Question answering systems: Survey and trends. *Procedia Computer Science*, 73:366–375.
- Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. **Generating sentences from a continuous space**. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.
- Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. **Extracting social power relationships from natural language**. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 773–782, Stroudsburg, PA, USA. Association for Computational Linguistics.
- L Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. 2020. Scalable and explainable legal prediction. *Artificial Intelligence and Law*, pages 1–26.
- Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. **Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

- Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.
- M. D. Bruyn, E. Lotfi, Jeska Buhmann, and W. Daelemans. 2020. Bart for knowledge grounded conversations. In *Converse@KDD*.
- John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1301–1309. Association for Computational Linguistics.
- Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*, pages 681–684.
- Yu Cao, Wei Bi, Meng Fang, and Dacheng Tao. 2020. [Pretrained language models for dialogue generation with multiple input sources](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 909–917, Online. Association for Computational Linguistics.
- Dallas Card and Noah A. Smith. 2020. [On consequentialism and fairness](#). *Frontiers in Artificial Intelligence*, 3.
- Jordan Carpenter, Daniel Preotiuc-Pietro, Lucie Flekova, Salvatore Giorgi, Courtney Hagan, Margaret L Kern, Anneke EK Buffone, Lyle Ungar, and Martin EP Seligman. 2016. Real men don’t say “cute” using automatic language analysis to isolate inaccurate aspects of stereotypes. *Social Psychological and Personality Science*.
- Amanda Cercas Curry and Verena Rieser. 2018. [#MeToo Alexa: How conversational systems respond to sexual harassment](#). In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 7–14, New Orleans, Louisiana, USA. Association for Computational Linguistics.
- Khyathi Chandu, Eric Nyberg, and Alan W Black. 2019a. [Storyboarding of recipes: Grounded contextual generation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6040–6046, Florence, Italy. Association for Computational Linguistics.
- Khyathi Chandu, Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. 2019b. “my way of telling a story”: Persona based grounded story generation. In *Proceedings of the Second Workshop on Storytelling*, pages 11–21.
- Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. Neural sentence ordering. *arXiv preprint arXiv:1607.06952*.
- Colin Cherry, Greg Durrett, George Foster, Reza Haffari, Shahram Khadivi, Nanyun Peng, Xiang Ren, and Swabha Swayamdipta, editors. 2019. *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Association for Computational Linguistics, Hong Kong, China.

- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.
- Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.
- Jennifer Coates. 2015. *Women, men and language: A sociolinguistic account of gender differences in language*. Routledge.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.
- Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM.
- Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.
- Nikolas Coupland. 2007. *Style: Language Variation and Identity*. Key Topics in Sociolinguistics. Cambridge University Press.
- Kate Crawford. 2017. The trouble with bias, 2017. URL <http://blog.revolutionanalytics.com/2017/12/the-trouble-with-bias-by-kate-crawford.html>. Invited Talk by Kate Crawford at NIPS.
- Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. 2018. Deep attentive sentence ordering network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4340–4349.
- Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. Association for Computational Linguistics.
- Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. [A computational approach to politeness with application to social factors](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.
- Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 update summarization task. In *In TAC 2008 Workshop - Notebook papers and results*, pages 10–23.



- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: a simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.
- Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. [Racial bias in hate speech and abusive language detection datasets](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.
- Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Chrysanne DiMarco. 1994. Stylistic choice in machine translation. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.
- Chrysanne DiMarco and Graeme Hirst. 1990. [Accounting for style in machine translation](#). In *Proceedings of the Third International Conference on Theoretical Issues in Machine Translation*, Austin.
- Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. [Queens are powerful too: Mitigating gender bias in dialogue generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020b. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.
- Xiyu Ding, Michael Barnett, Ateev Mehrotra, and Timothy Miller. 2020. [Methods for extracting information from messages from primary care providers to specialists](#). In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 1–6, Online. Association for Computational Linguistics.
- Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886. Association for Computational Linguistics.

- Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. "hey google is it ok if i eat you?": Initial explorations in child-agent interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children, IDC '17*, page 595–600, New York, NY, USA. Association for Computing Machinery.
- Duolingo. 2011. duolingo. <https://www.duolingo.com/> (accessed April 9, 2021).
- Nouha Dziri, Ehsan Kamaloo, Kory W Mathewson, and Osmar Zaiane. 2018. Augmenting neural response generation with context-aware topical attention. *arXiv preprint arXiv:1811.01063*.
- Penelope Eckert. 2019. The limits of meaning: Social indexicality, variation, and the cline of interiority. *Language*, 95(4):751–776.
- Penelope Eckert and Sally McConnell-Ginet. 2003. *Language and gender*. Cambridge University Press.
- Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. **Frames: a corpus for adding memory to goal-oriented dialogue systems**. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.
- Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. **Generating medical reports from patient-doctor conversations using sequence-to-sequence models**. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.
- Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. 2020. Text editing by command. *arXiv preprint arXiv:2010.12826*.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. *arXiv preprint arXiv:1902.01109*.
- Jessica Fidler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104.
- Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580.
- James Fieser. 1995. Ethics. <https://iep.utm.edu/ethics/> (accessed: 11-03-2020).

- Seeger Fisher and Brian Roark. 2008. Query-focused supervised sentence ranking for update summaries. In *TAC*.
- Susan T Fiske. 1993. Controlling other people: The impact of power on stereotyping. *American psychologist*, 48(6):621.
- Lucie Flekova and Iryna Gurevych. 2013. Can we hide in the web? large scale simultaneous age and gender author profiling in social media. In *CLEF 2012 Labs and Workshop, Notebook Papers*. Citeseer.
- Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146.
- Juri Ganitkevitch and Chris Callison-Burch. 2014. [The multilingual paraphrase database](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4276–4283, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Juri Ganitkevitch, Chris Callison-Burch, Courtney Napoles, and Benjamin Van Durme. 2011. [Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1168–1179, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. [The WebNLG challenge: Generating text from RDF data](#). In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Marta Gattius. 2017. [Personalized questions, answers and grammars: Aiding the search for relevant web information](#). In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 203–207, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. [Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

- Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *ACL*, volume 1, pages 634–642.
- Daniel Gildea and Daniel Jurafsky. 2000. [Automatic labeling of semantic roles](#). In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong. Association for Computational Linguistics.
- Laurent Gillard, Patrice Bellot, and Marc El-Bèze. 2006. Question answering evaluation survey. In *LREC*, pages 1133–1138. Citeseer.
- Andreea Godea and Rodney Nielsen. 2018. [Annotating educational questions for student response analysis](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. [Reinforcement learning based text style transfer without parallel training corpus](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3168–3180, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jingjing Gong, Xinchu Chen, Xipeng Qiu, and Xuanjing Huang. 2016. End-to-end neural sentence ordering using pointer network. *arXiv preprint arXiv:1611.04953*.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680.
- David Graff and Christopher Cieri. 2003. English Gigaword LDC2003T05. In *Philadelphia: Linguistic Data Consortium*.
- Inc Grammarly. 2009. grammarly. <https://www.grammarly.com/> (accessed April 9, 2021).
- Nicholas Greenquist, Doruk Kilitcioglu, and Anasse Bari. 2019. Gkb: A predictive analytics framework to generate online product recommendations. In *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*, pages 414–419. IEEE.
- Junliang Guo, Xu Tan, Di He, Tao Qin, Linli Xu, and Tie-Yan Liu. 2019. Non-autoregressive neural machine translation with enhanced decoder input. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3723–3730.

- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.
- Jiale Han, Bo Cheng, and Xizhou Wang. 2020. Two-phase hypergraph based reasoning with dynamic relations for multi-hop kbqa. In *IJCAI*.
- Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. [Towards a critical race methodology in algorithmic fairness](#). In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT\* '20*, page 501–512, New York, NY, USA. Association for Computing Machinery.
- Vrindavan Harrison, Lena Reed, Shereen Oraby, and Marilyn Walker. 2019. Maximizing stylistic control and semantic accuracy in nlg: Personality variation and discourse contrast. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*, pages 1–12.
- Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. [A probabilistic formulation of unsupervised text style transfer](#). In *International Conference on Learning Representations*.
- Barbara Heinisch and Vesna Lušicky. 2019. [User expectations towards machine translation: A case study](#). In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 42–48, Dublin, Ireland. European Association for Machine Translation.
- Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129.
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- E.T. Higgins and G.R. Semin. 2001. [Communication and social psychology](#). In Neil J. Smelser and Paul B. Baltes, editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 2296 – 2299. Pergamon, Oxford.
- Cong Duy Vu Hoang, Trevor Cohn, and Gholamreza Haffari. 2016. Incorporating side information into recurrent neural network language models. In *Proceedings of the 2016 Conference*

- of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1255.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *International Conference on Learning Representations*.
- Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- John Hooker. 2018. *Taking Ethics Seriously: Why Ethics Is an Essential Tool for the Modern Workplace*. Taylor and Francis.
- John N. Hooker and Tae Wan N. Kim. 2018. [Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic](#). In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18*, page 130–136, New York, NY, USA. Association for Computing Machinery.
- Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 752–762.
- Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. [“you sound just like your father” commercial machine translation systems include stylistic biases](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.
- Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, pages 452–461. International World Wide Web Conferences Steering Committee.
- Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 483–488.
- Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 591–598.

- Ed Hovy. 1993. From interclausal relations to discourse structure - a long way behind, a long way ahead. In Helmut Horacek and Michael Zock, editors, *New Concepts in Natural Language Generation: Planning, Realization and Systems*. St. Martin's Press, Inc., USA.
- Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.
- Eduard H. Hovy. 1988. [Planning coherent multisentential text](#). In *26th Annual Meeting of the Association for Computational Linguistics*, pages 163–169, Buffalo, New York, USA. Association for Computational Linguistics.
- Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. 2020a. [What makes a good story? designing composite rewards for visual storytelling](#). In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, New York, USA.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020b. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.
- Zhiting Hu, Haoran Shi, Bowen Tan, Wentao Wang, Zichao Yang, Tiancheng Zhao, Junxian He, Lianhui Qin, Di Wang, Xuezhe Ma, et al. 2019. Texar: A modularized, versatile, and extensible toolkit for text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 159–164.
- Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. [Toward controlled generation of text](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596, International Convention Centre, Sydney, Australia. PMLR.
- Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P. Xing. 2018. [On unifying deep generative models](#). In *International Conference on Learning Representations*.
- Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019a. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.
- Qiuyuan Huang, Zhe Gan, Asli Celikyilmaz, Dapeng Wu, Jianfeng Wang, and Xiaodong He. 2019b. Hierarchically structured reinforcement learning for topically coherent visual story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8465–8472.
- Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.
- Alan Jaffe. 2017. [Generating image descriptions using multilingual data](#). In *Proceedings of the Second Conference on Machine Translation*, pages 458–464, Copenhagen, Denmark. Association for Computational Linguistics.

- Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.
- Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. [WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Eric Jardine. 2016. Tor, what is it good for? political repression and the use of online anonymity-granting technologies. *New Media & Society*.
- Jianshu Ji, Qinlong Wang, Kristina Toutanova, Yongen Gong, Steven Truong, and Jianfeng Gao. 2017. A nested attention neural hybrid model for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 753–762, Vancouver, Canada.
- Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. [Disentangled representation learning for non-parallel text style transfer](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.
- Robert Johnson and Adam Cureton. 2019. Kant’s Moral Philosophy. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, spring 2019 edition. Metaphysics Research Lab, Stanford University.
- Kristiina Jokinen, Hideki Tanaka, and Akio Yokoo. 1998. [Context management with topics for spoken dialogue systems](#). In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 631–637, Montreal, Quebec, Canada. Association for Computational Linguistics.
- Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proc. of the Workshop on Noisy User-generated Text*, pages 9–18.
- Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. [MS-UEdin submission to the WMT2018 APE shared task: Dual-source transformer for automatic post-editing](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 822–826, Belgium, Brussels. Association for Computational Linguistics.
- Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO.
- Peter H. Kahn Jr, Heather E. Gary, and Solace Shen. 2013. [Children’s social relationships with current and near-future robots](#). *Child Development Perspectives*, 7(1):32–37.



- Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (male, bachelor) and (female, Ph.D) have different connotations: Parallely annotated stylistic language dataset with multiple personas. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1696–1706, Hong Kong, China. Association for Computational Linguistics.
- Dongyeop Kang and Eduard Hovy. 2019. xslue: A benchmark and analysis platform for cross-style language understanding and evaluation. In <https://arxiv.org>.
- Immanuel Kant. 1785. *Groundwork for the Metaphysics of Morals*. Yale University Press.
- Geoff F Kaufman and Lisa K Libby. 2012. Changing beliefs and behavior through experience-taking. *Journal of personality and social psychology*, 103(1):1.
- Shari Kendall, Deborah Tannen, et al. 1997. Gender and language in the workplace. *Gender and Discourse*. London: Sage, pages 81–105.
- Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339, Austin, Texas. Association for Computational Linguistics.
- Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas.
- Seokhwan Kim, Rafael E. Banchs, and Haizhou Li. 2015. Towards improving dialogue topic tracking performances with wikification of concept mentions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 124–128, Prague, Czech Republic. Association for Computational Linguistics.
- Gary King, Jennifer Pan, and Margaret E Roberts. 2017. How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American political science review*, 111(3):484–501.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proc. ICLR*.
- John Kleinig. 2009. *The Nature of Consent* \*.
- Bryan Klimt and Yiming Yang. 2004. Introducing the enron corpus. In *CEAS*.
- Donald Ervin Knuth. 1998. *The art of computer programming, , Volume III, 2nd Edition*. Addison-Wesley.

- Ioannis Konstas and Mirella Lapata. 2012. Concept-to-text generation via discriminative reranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 369–378. Association for Computational Linguistics.
- Iliia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. [Importance of search and evaluation strategies in neural dialogue modeling](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.
- Rohit Kumar, Carolyn P Rosé, Yi-Chia Wang, Mahesh Joshi, and Allen Robinson. 2007. Tutorial dialogue as adaptive collaborative learning support. In *Proceedings of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, pages 383–390.
- Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. [Topics to avoid: Demoting latent confounds in text classification](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4153–4163, Hong Kong, China. Association for Computational Linguistics.
- Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.
- Matt J Kusner and José Miguel Hernández-Lobato. 2016. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*.
- Robin Tolmach Lakoff and Mary Bucholtz. 2004. *Language and woman’s place: Text and commentaries*, volume 3. Oxford University Press, USA.
- Alex M Lamb, Anirudh Goyal Alias Parth Goyal, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. In *Advances in neural information processing systems*, pages 4601–4609.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.
- Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 545–552. Association for Computational Linguistics.

- Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.
- Brian Larson. 2017. [Gender as a variable in natural-language processing: Ethical considerations](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.
- Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. *arXiv preprint arXiv:2004.05569*.
- Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213. Association for Computational Linguistics.
- Jochen L. Leidner and Vassilis Plachouras. 2017. [Ethical by design: Ethics best practices for natural language processing](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.
- Dave Lewis, Joss Moorkens, and Kaniz Fatema. 2017. [Integrating the management of personal data protection and open science with research ethics](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 60–65, Valencia, Spain. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.
- Paul Lewis and Paul Hilder. 2018. Leaked: Cambridge analytica's blueprint for trump victory. <https://www.theguardian.com/uk-news/2018/mar/23/leaked-cambridge-analyticas-blueprint-for-trump-victory> (accessed April 9, 2021).
- Chen Li, Yang Liu, and Lin Zhao. 2015. Improving update summarization via supervised ILP and sentence reranking. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1317–1322.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics:*

- Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.
- Jiwei Li and Eduard Hovy. 2014. A model of coherence based on distributed sentence representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2039–2048.
- Jiwei Li and Dan Jurafsky. 2017. Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209.
- Juncen Li, Robin Jia, He He, and Percy Liang. 2018a. [Delete, retrieve, generate: a simple approach to sentiment and style transfer](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.
- Wei Li, Xinyan Xiao, Yajuan Lyu, and Yuanzhuo Wang. 2018b. [Improving neural abstractive document summarization with structural regularization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4078–4087, Brussels, Belgium. Association for Computational Linguistics.
- Xiang Lisa Li and Jason Eisner. 2019. Specializing word embeddings (for parsing) by information bottleneck. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2744–2754.
- Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4*, pages 45–51. Association for Computational Linguistics.
- Crisis Text Line. 2013. Crisis text line. <https://www.crisistextline.org/> (accessed April 9, 2021).
- Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.
- Chao-Hong Liu, editor. 2018. *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*. Association for Machine Translation in the Americas, Boston, MA.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

- Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. [Generating wikipedia by summarizing long sequences](#). In *International Conference on Learning Representations*.
- Yang Liu and Mirella Lapata. 2018. Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75.
- Ye Liu, Shaika Chowdhury, Chenwei Zhang, Cornelia Caragea, and Philip S. Yu. 2020. [Interpretable multi-step reasoning with knowledge extraction on complex healthcare question answering](#).
- Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018a. Content preserving text generation with attribute controls. In *Advances in Neural Information Processing Systems*, pages 5103–5113.
- Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2018b. Sentence ordering and coherence modeling using recurrent neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. [The many dimensions of algorithmic fairness in educational applications](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10, Florence, Italy. Association for Computational Linguistics.
- Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.
- Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. 2017. Deep photo style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4990–4998.
- Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015a. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.
- Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015b. Effective approaches to attention-based neural machine translation. In *Proc. EMNLP*.
- Kelvin Luu, Chenhao Tan, and Noah A Smith. 2019. Measuring online debaters’ persuasive skill from text over time. *Transactions of the Association for Computational Linguistics*, 7:537–550.
- Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. [PowerTransformer: Unsupervised controllable revision for biased language correction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

- Aman Madaan, Dheeraj Rajagopal, Yiming Yang, Abhilasha Ravichander, Eduard Hovy, and Shrimai Prabhunoye. 2020a. [Eigen: Event influence generation using pre-trained language models](#).
- Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhunoye. 2020b. [Politeness transfer: A tag and generate approach](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.
- Nitin Madnani and Bonnie J Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.
- Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2019. [Common-sense knowledge base completion with structural and semantic context](#).
- Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proce. EACL*, volume 1, pages 881–893.
- William Mann and Sandra Thompson. 1988. [Rhetorical structure theory: Toward a functional theory of text organization](#). *Text*, 8:243–281.
- Daniel Marcu. 1997. From local to global coherence: A bottom-up approach to text planning. In *AAAI/IAAI*.
- Elijah Mayfield and Alan W Black. 2020. [Should you fine-tune BERT for automated essay scoring?](#) In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA. Online. Association for Computational Linguistics.
- Elijah Mayfield, Michael Madaio, Shrimai Prabhunoye, David Gerritsen, Brittany McLaughlin, Ezekiel Dixon-Román, and Alan W Black. 2019. [Equity beyond bias in language technologies for education](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 444–460, Florence, Italy. Association for Computational Linguistics.
- Stephen Mayhew, Klinton Bicknell, Chris Brust, Bill McDowell, Will Monroe, and Burr Settles. 2020. Simultaneous translation and paraphrase for language education. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 232–243.
- Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. 2007. [Topic and role discovery in social networks with experiments on enron and academic email](#). *J. Artif. Int. Res.*, 30(1):249–272.
- T. McEnery. 2005. [Swearing in english: Bad language, purity and power from 1586 to the present](#). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*, pages 1–248.

- Hongyuan Mei, TTI UChicago, Mohit Bansal, and Matthew R Walter. 2016. What to talk about and how? selective generation using LSTMs with coarse-to-fine alignment. In *Proceedings of NAACL-HLT*, pages 720–730.
- Paul Michel and Graham Neubig. 2018. [Extreme adaptation for personalized neural machine translation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.
- Corporation Microsoft. 2021. Microsoft editor. bring out your best writing anywhere you write. <https://www.microsoft.com/en-us/microsoft-365/microsoft-editor> (accessed April 9, 2021).
- Margot Mieskes. 2017. [A quantitative study of data in the NLP community](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 23–29, Valencia, Spain. Association for Computational Linguistics.
- Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55.
- Hideki Mima, O. Furuse, and H. Iida. 1997. Improving performance of transfer-driven machine translation with extra-linguistic information from context, situation and environment. In *IJCAI*.
- Shachar Mirkin and Jean-Luc Meunier. 2015. [Personalized machine translation: Predicting translational preferences](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbon, Portugal. Association for Computational Linguistics.
- Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. [Motivating personality-aware machine translation](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108, Lisbon, Portugal. Association for Computational Linguistics.
- Margaret Mitchell, Kathleen McCoy, David McDonald, and Aoife Cahill, editors. 2014. *Proceedings of the 8th International Natural Language Generation Conference (INLG)*. Association for Computational Linguistics, Philadelphia, Pennsylvania, U.S.A.
- Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*.
- Evgeny Morozov. 2013. *To save everything, click here: The folly of technological solutionism*. Public Affairs.
- Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. [Image-grounded conversations: Multimodal context for natural question and response generation](#). In *Proceedings of the Eighth International*

- Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California.
- Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Nobushige Doi, Yusuke Oda, Ondřej Bojar, Shantipriya Parida, Isao Goto, and Hidaya Mino, editors. 2019. *Proceedings of the 6th Workshop on Asian Translation*. Association for Computational Linguistics, Hong Kong, China.
- Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290. Association for Computational Linguistics.
- Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. Technical report, Microsoft Research.
- Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593.
- Tong Niu and Mohit Bansal. 2018a. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.
- Tong Niu and Mohit Bansal. 2018b. [Polite dialogue generation without parallel data](#). *Transactions of the Association for Computational Linguistics*, 6:373–389.
- Xing Niu and Marine Carpuat. 2019. Controlling neural machine translation formality with synthetic supervision. *arXiv preprint arXiv:1911.08706*.
- Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819.
- Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1659–1666.



- Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alexander Fraser, Shankar Kumar, Libin Shen, David A Smith, Katherine Eng, et al. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 161–168.
- Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. [Multimodal abstractive summarization for how2 videos](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596, Florence, Italy. Association for Computational Linguistics.
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Tanmay Parekh, Emily Ahn, Yulia Tsvetkov, and Alan W Black. 2020. [Understanding linguistic accommodation in code-switched human-machine dialogues](#). In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 565–577, Online. Association for Computational Linguistics.
- Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. [Ethical considerations in NLP shared tasks](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 66–73, Valencia, Spain. Association for Computational Linguistics.
- Romain Paulus, Caiming Xiong, and Richard Socher. 2018. [A deep reinforced model for abstractive summarization](#). In *International Conference on Learning Representations*.
- Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, pages 37–44. ACM.
- Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49.

- Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. [Email formality in the workplace: A case study on the Enron corpus](#). In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 86–95, Portland, Oregon. Association for Computational Linguistics.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes, editors. 2020. *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1st virtual meeting.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. [Hypothesis only baselines in natural language inference](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.
- John Pougué Biyong, Bo Wang, Terry Lyons, and Alejo Nevado-Holgado. 2020. [Information extraction from Swedish medical prescriptions with sig-transformer encoder](#). In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 41–54, Online. Association for Computational Linguistics.
- Vinodkumar Prabhakaran, Camilla Griffiths, Hang Su, Prateek Verma, Nelson Morgan, Jennifer Eberhardt, and Dan Jurafsky. 2018. Detecting institutional dialog acts in police traffic stops. *Transactions of the Association for Computational Linguistics*, 6:467–481.
- Vinodkumar Prabhakaran, Emily E. Reid, and Owen Rambow. 2014. [Gender and power: How gender and gender environment affect manifestations of power](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1965–1976, Doha, Qatar. Association for Computational Linguistics.
- Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020a. [Exploring controllable text generation techniques](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Shrimai Prabhumoye, Margaret Li, Jack Urbanek, Emily Dinan, Douwe Kiela, Jason Weston, and Arthur Szlam. 2020b. I love your chain mail! making knights smile in a fantasy game world: Open-domain goal-oriented dialogue agents. *arXiv preprint arXiv:2002.02878*.
- Shrimai Prabhumoye, Elijah Mayfield, and Alan W Black. 2019a. [Principled frameworks for evaluating ethics in nlp systems](#).

- Shrimai Prabhumoye, Chris Quirk, and Michel Galley. 2019b. [Towards content transfer through grounded text generation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2622–2632, Minneapolis, Minnesota. Association for Computational Linguistics.
- Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. 2020c. [Topological sort for sentence ordering](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2783–2792, Online. Association for Computational Linguistics.
- Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. [Style transfer through back-translation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.
- Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. [Learning to deceive with attention-based explanations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.
- Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.
- J. Pugh. 2020. *Autonomy, Rationality, and Contemporary Bioethics [Internet]*. Oxford University Press, Oxford (UK).
- Srikrishna Raamadhurai, Ryan Baker, and Vikraman Poduval. 2019. [Curio SmartChat : A system for natural language question answering for self-paced k-12 learning](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 336–342, Florence, Italy. Association for Computational Linguistics.
- Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. [Personalized machine translation: Preserving original author traits](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

- Dheeraj Rajagopal, Niket Tandon, Peter Clarke, Bhavana Dalvi, and Eduard Hovy. 2020. What-if i ask you to explain: Explaining the effects of perturbations in procedural text. *arXiv preprint arXiv:2005.01526*.
- Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.
- Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44.
- Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 129–140.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.
- Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. [Question answering for privacy policies: Combining computational and legal perspectives](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958, Hong Kong, China. Association for Computational Linguistics.
- Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. [CoQA: A conversational question answering challenge](#). *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Sravana Reddy and Kevin Knight. 2016. [Obfuscating gender in social media writing](#). In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26, Austin, Texas. Association for Computational Linguistics.
- Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge University Press, USA.
- Michaela Regneri, Marcus Rohrbach, Dominikus Wetzels, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. [Grounding action descriptions in videos](#). *Transactions of the Association for Computational Linguistics*, 1:25–36.
- Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, USA.
- Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2019. [Adversarial decomposition of text representation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

- Volume 1 (Long and Short Papers)*, pages 815–825, Minneapolis, Minnesota. Association for Computational Linguistics.
- Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International journal of computer-supported collaborative learning*, 3(3):237–271.
- Carolyn Penstein Rosé, Ryan Carlson, Diyi Yang, Miaomiao Wen, Lauren Resnick, Pam Goldman, and Jennifer Sherer. 2014. Social factors that contribute to attrition in moocs. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 197–198.
- Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. *Automatic Keyword Extraction from Individual Documents*.
- W. D. Ross. 1930. *The Right and the Good*. Clarendon Press, Oxford (UK).
- Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proc. of the First Workshop on Ethics in Natural Language Processing*, page 74.
- Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, editors. 2020. *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Online.
- Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. [A neural attention model for abstractive sentence summarization](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.
- Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2018. [Commonsense knowledge base completion and generation](#). In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 141–150, Brussels, Belgium. Association for Computational Linguistics.
- Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. 2019. [Integrating ethics within machine learning courses](#). *ACM Trans. Comput. Educ.*, 19(4).
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. [The risk of racial bias in hate speech detection](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. [Social bias frames: Reasoning about social and power implications of language](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

- Mourad Sarrouiti and Said Ouatic El Alaoui. 2017. [A biomedical question answering system in BioASQ 2017](#). In *BioNLP 2017*, pages 296–301, Vancouver, Canada,. Association for Computational Linguistics.
- Allen Schmalz. 2018. [On the utility of lay summaries and AI safety disclosures: Toward robust, open research oversight](#). In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 1–6, New Orleans, Louisiana, USA. Association for Computational Linguistics.
- Tyler Schnoebelen. 2017. [Goal-oriented design for ethical machine learning and NLP](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 88–93, Valencia, Spain. Association for Computational Linguistics.
- Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.
- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. [MLSUM: The multilingual summarization corpus](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.
- Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. [What makes a good conversation? how controllable attributes affect human judgments](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.
- Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. 2019. Beyond dyadic interactions: Considering chatbots as community members. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Controlling politeness in neural machine translation via side constraints](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.
- Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.
- Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, 9(1):1–49.

- Sofia Serrano and Noah A. Smith. 2019. [Is attention interpretable?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.
- Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. [Predictive biases in natural language processing models: A conceptual framework and overview.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.
- Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. [Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation.](#) *CoRR*, abs/1706.09799.
- Solace Shen. 2015. *Children’s Conceptions of the Moral Standing of a Humanoid Robot of the Here and Now*. Ph.D. thesis.
- Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841.
- Jitesh Shetty and Jafar Adibi. 2004. The enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California*, 4(1):120–128.
- Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. [Engaging image chat: Modeling personality in grounded dialogue.](#) *CoRR*, abs/1811.00945.
- Yan Shvartzshanider, Ananth Balashankar, Thomas Wies, and Lakshminarayanan Subramanian. 2018. [RECIPE: Applying open domain question answering to privacy policies.](#) In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 71–77, Melbourne, Australia. Association for Computational Linguistics.
- Chunjin Song, Zhijie Wu, Yang Zhou, Minglun Gong, and Hui Huang. 2019. [Etnet: Error transition network for arbitrary style transfer.](#) In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Steven J. Spencer, Claude M. Steele, and Diane M. Quinn. 1999. Stereotype Threat and Women’s Math Performance. *Journal of Experimental Social Psychology*, 35:4–28.
- Andreas Stolcke. 2002. Srilm—an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.
- Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. [“transforming” delete, retrieve, generate approach for controlled text style transfer.](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.

- Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. [Mitigating gender bias in natural language processing: Literature review](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.
- Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. [A short review of ethical challenges in clinical natural language processing](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 80–87, Valencia, Spain. Association for Computational Linguistics.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.
- Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.
- Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. From neural sentence summarization to headline generation: a coarse-to-fine approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4109–4115.
- Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634.
- Deborah Tannen. 1991. *You just don't understand: Women and men in conversation*. Virago London.
- Deborah Tannen. 1993. *Gender and conversational interaction*. Oxford University Press.
- Robert Endre Tarjan. 1976. [Edge-disjoint spanning trees and depth-first search](#). *Acta Informatica*, 6(2):171–185.
- Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59.
- Youzhi Tian, Weiyan Shi, Chen Li, and Zhou Yu. 2020. [Understanding user resistance strategies in persuasive conversations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4794–4798, Online. Association for Computational Linguistics.
- Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683.



- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- S Verberne, LWJ Boves, NHJ Oostdijk, and PAJM Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, pages 735–737. Amsterdam: Association for Computing Machinery.
- Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proc. ICML Deep Learning Workshop*.
- Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. [RtGender: A corpus for studying differential responses to gender](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, et al. Multi-simlex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity. *Computational Linguistics*, pages 1–51.
- Ke Wang, Hang Hua, and Xiaojun Wan. 2019a. Controllable unsupervised text attribute transfer via editing entangled latent representation. In *Advances in Neural Information Processing Systems*, pages 11034–11044.
- Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019b. [Topic-guided variational auto-encoder for text generation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 166–177, Minneapolis, Minnesota. Association for Computational Linguistics.
- Xu Wang, Diyi Yang, Miaomiao Wen, Kenneth Koedinger, and Carolyn P Rosé. 2015. Investigating how student’s cognitive behavior in mooc discussion forums affect learning gains. *International Educational Data Mining Society*.
- Xuewei Wang, Weiyang Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019c. Persuasion for good: Towards a personalized persuasive dialogue system for

- social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.
- Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.
- Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6383–6389.
- Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. [AirDialogue: An environment for goal-oriented dialogue research](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3844–3854, Brussels, Belgium. Association for Computational Linguistics.
- Sean Welleck, Kianté Brantley, Hal Daumé Iii, and Kyunghyun Cho. 2019. Non-monotonic sequential text generation. In *International Conference on Machine Learning*, pages 6716–6726.
- Sean Welleck, Ilya Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. [Neural text generation with unlikelihood training](#). In *International Conference on Learning Representations*.
- Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.
- John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 274–285.
- Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.
- Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.
- Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.
- Florian Wolf and Edward Gibson. 2006. *Coherence in natural language: data structures and applications*. MIT Press.

- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. [Mask and infill: Applying masked language model for sentiment transfer](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5271–5277. International Joint Conferences on Artificial Intelligence Organization.
- Ziang Xie. 2017. Neural text generation: A practical guide. *arXiv preprint arXiv:1711.09534*.
- Peng Xu, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019. [Unsupervised controllable text generation with global variation discovery and disentanglement](#). *CoRR*, abs/1905.11975.
- Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. [Controlling the voice of a sentence in Japanese-to-English neural machine translation](#). In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 203–210, Osaka, Japan. The COLING 2016 Organizing Committee.
- Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. 2013. ” turn on, tune in, drop out”: Anticipating student dropouts in massive open online courses. In *Neural Information Processing Systems: Workshop on Data-Driven Education (NIPS 2013)*.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018a. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018b. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems*, pages 7287–7298.
- Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.
- Jen-Yuan Yeh and Aaron Harnly. 2006. Email thread reassembly using similarity matching. In *Conference on Email and Anti-Spam*. Conference on Email and Anti-Spam.
- Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-first AAAI conference on artificial intelligence*.

- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending against neural fake news. *Neurips*.
- Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5276–5289.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. [Gender bias in contextualized word embeddings](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. [Men also like shopping: Reducing gender bias amplification using corpus-level constraints](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.
- Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. [Low-resource knowledge-grounded dialogue generation](#). In *International Conference on Learning Representations*.
- Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. [Knowledge-grounded dialogue generation with pre-trained language models](#).
- Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. [A dataset for document grounded conversations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.
- Yiheng Zhou, Yulia Tsvetkov, Alan W Black, and Zhou Yu. 2020. [Augmenting non-collaborative dialog systems with explicit semantic and strategic dialog history](#). In *International Conference on Learning Representations*.
- J. Zhu, T. Park, P. Isola, and A. A. Efros. 2017. [Unpaired image-to-image translation using cycle-consistent adversarial networks](#). In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251.
- Pierre Zweigenbaum. 2009. [Knowledge and reasoning for medical question-answering](#). In *Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions (KRAQ 2009)*, pages 1–2, Suntec, Singapore. Association for Computational Linguistics.