

## ABSTRACT

Title of dissertation:      **FORMALITY STYLE TRANSFER  
WITHIN AND ACROSS LANGUAGES  
WITH LIMITED SUPERVISION**

Xing Niu, Doctor of Philosophy, 2019

Dissertation directed by:   **Marine Carpuat  
Department of Computer Science**

While much natural language processing work focuses on analyzing language content, language style also conveys important information about the situational context and purpose of communication. When editing an article, professional editors take into account the target audience to select appropriate word choice and grammar. Similarly, professional translators translate documents for a specific audience and often ask what is the expected tone of the content when taking a translation job.

Computational models of natural language should consider both their meaning and style. Controlling style is an emerging research area in text rewriting and is under-investigated in machine translation. In this dissertation, we present a new perspective which closely connects formality transfer and machine translation: we aim to control style in language generation with a focus on rewriting English or translating French to English with a desired formality. These are challenging tasks because annotated examples of style transfer are only available in limited quantities.

We first address this problem by inducing a lexical formality model based on word embeddings and a small number of representative formal and informal words.

This enables us to assign sentential formality scores and rerank translation hypotheses whose formality scores are closer to user-provided formality level. To capture broader formality changes, we then turn to neural sequence to sequence models. Joint modeling of formality transfer and machine translation enables formality control in machine translation without dedicated training examples. Along the way, we also improve low-resource neural machine translation.

FORMALITY STYLE TRANSFER WITHIN AND ACROSS LANGUAGES  
WITH LIMITED SUPERVISION

by

Xing Niu

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2019

Advisory Committee:

Dr. Marine Carpuat, Chair/Advisor

Dr. Douglas Oard, Dean's Representative

Dr. Jordan Boyd-Graber, Member

Dr. Furong Huang, Member

Dr. Philipp Koehn, Member

© Copyright by  
Xing Niu  
2019

## Acknowledgments

I would like to express my gratitude to all the people who made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost, I would like to thank my advisor Marine Carpuat, who introduced me to the realm of machine translation that cultivated my passion. She offered me great freedom as well as insightful thoughts to let me pursue my ideas and explore new research directions. She worked closely with me for all deadlines and significantly improved all of my writing pieces and presentations with her suggestions and edits. She also created a great environment in which I could spend most of my time on research and effectively collaborate with others. Most of all, her scrupulous attitude influenced me profoundly, and I have been benefited tremendously from this value. I feel very fortunate to work with Marine.

I would also like to thank the other members of my committee, Douglas Oard, Philipp Koehn, Jordan Boyd-Graber, and Furong Huang, for their insightful comments and questions that helped me improve this dissertation.

I am also grateful to my colleagues and friends in the Computational Linguistics and Information Processing Lab. I received valuable advice from Jimmy Lin, Philip Resnik, Hal Daum III, and Naomi Feldman. I enjoyed working together with my excellent collaborators Jinfeng Rao, Yogarshi Vyas, Weijia Xu, Sudha Rao, Marianna Martindale, and Amittai Axelrod; as well as the MATERIAL teammates Petra Galuscakova, Suraj Nair, and Elena Zotkina. I appreciate generous help from and

creative discussion with my labmates Hua He, Fenfei Guo, Feng Shi, Weiwei Yang, Chen Zhao, Sweta Agrawal, Aquia Richburg, Eleftheria Briakou, Joseph Barrow, Han-Chin Shing, Khanh Nguyen, Ahmed Elgohary, Amr Sharaf, Kianté Brantley, Allyson Ettinger, Craig Thorburn, Denis Peskov, Pedro Rodriguez, He He, Lingzi Hong, Jiahui Wu, and Yulu Wang. I thank Joe Webster and Bahar Azami for their prompt technical support.

My graduate journey could not be complete without two wonderful summer internships. Jagadeesh Jagarlamudi at Google was outstandingly cooperative and patient with me. Michael Denkowski and Alon Lavie at Amazon helped me to be productive in emerging research topics.

Thanks to all my friends. I am fortunate to share time with many lovely people at the University of Maryland: Xingfeng He, Hao Zhou, Xiyang Dai, Zebao Gao, Zheng Xu, Nairui Zhou and Shuhua Zhu.

Last but not least, I owe my sincerest appreciations to my family members, especially my parents, and to my beloved, Yingqi Zhang. They are always with me.

## Table of Contents

Acknowledgements	ii
List of Tables	viii
List of Figures	xi
1 Introduction	1
1.1 Research Problems	3
1.2 Roadmap	5
1.2.1 Modeling Lexical Stylistic Variations	6
1.2.2 Controlling Formality in Phrase-Based MT	6
1.2.3 Low-Resource Neural Machine Translation	7
1.2.4 Joint Model of Neural Formality Transfer and FSMT	8
1.2.5 Neural FSMT with Synthetic Supervision	9
1.3 Contributions	10
2 Background	12
2.1 Machine Translation	12
2.1.1 Phrase-Based Machine Translation	13
2.1.2 Neural Machine Translation	15
2.1.3 Multilingual and Zero-Shot Neural Machine Translation	18
2.1.4 Round-Trip Neural Machine Translation	19
2.2 Stylistic Variations	21
2.2.1 Formality: Definition and Discussion	21
2.2.2 Modeling Stylistic Variations	22
2.3 Stylistic Variations in Language Generation	24
2.3.1 Conditional Language Generation	24
2.3.2 Textual Style Transfer or Rewriting	25
2.3.3 Controlling the Output Style in Machine Translation	28
2.3.4 Evaluation of Style-Constrained Language Generation	29

3	Discovering Lexical Stylistic Variations in Distributional Vector Space Models	31
3.1	Approach	32
3.2	Qualitative Analysis of Latent Style Dimensions	33
3.2.1	Models Settings	33
3.2.2	Analysis	34
3.3	Extrinsic Evaluation: Lexical Formality Scoring	36
3.3.1	Identifying a Style Subspace	37
3.3.1.1	Experimental Set-Up	37
3.3.1.2	Results	38
3.3.2	SVM-Based Ranking vs. Other Formality Models	39
3.3.2.1	Formality Models	40
3.3.2.2	Results	41
3.3.3	Error Analysis	42
3.4	Summary	44
4	Reranking-Based Formality-Sensitive Machine Translation	46
4.1	Formality Modeling	47
4.1.1	Lexical Formality	47
4.1.1.1	Models Based on Word Representations	47
4.1.1.2	Models Based on Word Statistics	48
4.1.2	From Word to Sentence Formality	49
4.1.3	Evaluation	50
4.2	Formality-Sensitive Machine Translation	52
4.2.1	Experimental Set-Up	52
4.2.2	Automatic Evaluation	54
4.2.3	Human Assessment	55
4.3	Summary	57
5	Bi-Directional Low-Resource Neural Machine Translation	59
5.1	Bi-Directional Models with Synthetic Parallel Data	60
5.1.1	Approach	62
5.1.1.1	Bi-Directional NMT with Synthetic Parallel Data	62
5.1.1.2	Monolingual Data Selection	63
5.1.2	Experiments	63
5.1.2.1	Data	64
5.1.2.2	NMT Configuration	65
5.1.2.3	Uni-Directional NMT	67
5.1.2.4	Bi-Directional NMT	68
5.1.2.5	Fine-Tuning and Re-Decoding	70
5.1.2.6	Size of Selected Monolingual Data	71
5.1.2.7	Domain Adaptation	72
5.2	Bi-Directional Differentiable Input Reconstruction	73
5.2.1	Approach	74
5.2.1.1	Bi-Directional Reconstruction	75
5.2.1.2	Differentiable Sampling	76

5.2.2	Experiments . . . . .	77
5.2.2.1	Tasks and Data . . . . .	77
5.2.2.2	Model Configuration and Baseline . . . . .	78
5.2.2.3	Contrastive Reconstruction Model . . . . .	79
5.2.2.4	Results . . . . .	80
5.3	Summary . . . . .	82
6	Multi-Task Neural Formality Transfer and FSMT . . . . .	84
6.1	Approach . . . . .	85
6.1.1	Bi-Directional Formality Transfer . . . . .	86
6.1.2	Formality-Sensitive Machine Translation with Side Constraints . . . . .	86
6.1.3	Multi-Task Learning . . . . .	87
6.2	Experimental Set-Up . . . . .	89
6.3	Evaluation Protocol . . . . .	91
6.3.1	Automatic Evaluation . . . . .	91
6.3.2	Human Evaluation . . . . .	91
6.4	Formality Transfer Experiments . . . . .	95
6.4.1	Baseline Models . . . . .	95
6.4.2	Our Models . . . . .	95
6.4.3	Results . . . . .	96
6.4.4	Qualitative Analysis . . . . .	99
6.5	Formality-Sensitive Machine Translation Experiments . . . . .	102
6.5.1	Models . . . . .	102
6.5.2	Results . . . . .	103
6.5.3	Qualitative Analysis . . . . .	104
6.6	Summary . . . . .	106
7	Neural FSMT with Synthetic Supervision . . . . .	108
7.1	Approach . . . . .	109
7.1.1	Controlling the Output Language Formality . . . . .	110
7.1.2	Synthetic Supervision . . . . .	111
7.1.2.1	Online Style Inference (OSI) . . . . .	111
7.1.2.2	Online Target Inference (OTI) . . . . .	112
7.2	Auxiliary English Formality Control Evaluation . . . . .	113
7.3	FSMT Evaluation Set-Up . . . . .	116
7.3.1	Tasks and Data . . . . .	117
7.3.2	Baseline Models . . . . .	118
7.3.3	Implementation Details . . . . .	119
7.4	Automatic Evaluation of FSMT . . . . .	120
7.4.1	Lessons from BLEU . . . . .	120
7.4.2	Quantifying Differences Between Formal and Informal Outputs . . . . .	121
7.5	Human Evaluation of FSMT . . . . .	124
7.6	Qualitative Analysis . . . . .	128
7.7	Summary . . . . .	131

8	Conclusion and Future Work	132
8.1	Summary	132
8.2	Future Work	134
8.2.1	Modeling Formality in the Neural Architecture	134
8.2.2	A Broader Range of Tasks	135
8.2.3	Challenges of Joint Training	136
8.2.4	Differentiable Sampling for Unsupervised and Semi-Supervised Training	137
	Bibliography	139

## List of Tables

3.1	Representative word pairs for top principal components (indexed by $k$ ) are listed for both blogs and news corpora. A mixed variation of formality and American-British English (grey-boxed) can be characterized by the first principal component, but the following principal components seem vaguer in terms of interpreting stylistic variations. . . . .	35
3.2	Top (mis-)predicted CTRW word pairs, where $s_i$ is the SVM (formality) score for word $w_i$ . $w_2$ is supposed to be more formal than $w_1$ . † This word is more frequent than the other in a pair according to the blogs corpus. (‡/ † †/ † † means at least 10/100/1000 times more.) . . . . .	43
4.1	Sentence-level formality quantifying evaluation (RMSE) among different models with different vector spaces. . . . .	51
4.2	Translation quality (BLEU scores) on informal/neutral/formal sentence sets given different desired formality levels ( $-0.4, 0.0, 0.4$ ). Best results with statistical significance are highlighted. . . . .	54
4.3	Examples of variant translations to the same French source segment using low/high output formality levels ( $-0.4/0.4$ ) as parameters. In general the variations lie on the direction of formality as expected, but occasionally translation errors occur. . . . .	57
5.1	Data sizes of training, development, test, sample and monolingual sets. Sample data serves as the in-domain seed for data selection. . . . .	66
5.2	BLEU scores for uni-directional models (ID=U- $k$ ) and bi-directional NMT models (ID=B- $k$ ) trained on different combinations of real and synthetic parallel data. Models in B-5 <i>f</i> are fine-tuned from base models in B-1. Best models in B-6 <i>f</i> are fine-tuned from precedent models in B-5 <i>f</i> and underscored synthetic data is re-decoded using precedent models. The highest score within each box is highlighted. . . . .	67
5.3	Number of checkpoints ( $=  \text{updates} /1000$ for TL/SW $\leftrightarrow$ EN or $ \text{updates} /10,000$ for DE $\leftrightarrow$ EN) used by various NMT models. Bi-directional models (with fine-tuning) reduce training time significantly. . . . .	70

5.4	BLEU scores for bi-directional NMT models on Bible data. Models in B-5 <i>f</i> are fine-tuned from baseline models in B-1. Highlighted best models in B-6 <i>f</i> are fine-tuned from precedent models in B-5 <i>f</i> and underscored synthetic data is re-decoded using precedent models. Baseline models are significantly improved in terms of BLEU. . . . .	72
5.5	Experiments are conducted on four low-resource language pairs, in both translation directions. . . . .	78
5.6	BLEU scores on eight translation directions. The numbers before and after ‘±’ are the mean and standard deviation over five randomly seeded models. Our proposed methods ( $\beta = 0/0.5$ ) achieve small but consistent improvements. $\Delta$ BLEU scores are in bold if mean–std is above zero while in red if the mean is below zero. . . . .	80
6.1	Automatic evaluation of Formality Transfer with BLEU scores. The bi-directional model with three stacked improvements achieves the best overall performance. The improvement over the second best system is statistically significant at $p < 0.05$ using bootstrap resampling (Koehn, 2004b). . . . .	96
6.2	Human evaluation of formality difference and meaning preservation. MultiTask-tag-style generates significantly more informal (F→I) English than NMT Combined ( $p < 0.05$ using the t-test, see Section 6.4.3). PBMT-random does not control formality effectively when comparing its informal (I) and formal (F) output (Section 6.5.2). Formality scores are relatively low because workers rarely choose “much more (in)formal”. All models preserve meaning equally well. . . . .	97
6.3	Sample model outputs for the Formality Transfer (FT) task. . . . .	100
6.4	BLEU scores of various FSMT models. “+Tag” indicates using formality tags for bilingual data. “Random” indicates using randomly selected bilingual data. . . . .	103
6.5	Sample model outputs for the Formality-Sensitive Machine Translation (FSMT) task. . . . .	105
7.1	BLEU scores for variants of side constraint in controlling style on all formality transfer and preservation directions. We report mean and standard deviation over five randomly seeded models. $\Delta$ BLEU between each model and the widely used TAG-SRC methods show that (1) blocking the visibility of source tags from the encoder (TAG-SRC-BLOCK) limits its formality control ability; (2) using style tags on both source and target sides (TAG-SRC-TGT) helps control formality better, especially for formality preservation tasks. . . . .	115
7.2	Statistics of French-English corpora. . . . .	118

7.3	All FSMT systems achieve better BLEU scores when the intended formality matches the nature of the text being translated (scores are grayed otherwise). LEPOD scores (all scores are percentages) show that synthetic supervision introduces more changes between formal and informal outputs than baselines, and Online Style Inference produces the most diverse informal/formal translations. . . . .	122
7.4	Types of the differences between informal and formal translations. Examples are drawn from the output of Online Style Inference. . . . .	129
7.5	Heuristic analysis of the differences between informal and formal translations. Both synthetic supervision methods introduce more changes between formal and informal translations. Online Target Inference usually performs simple substitutions while Online Style Inference performs more less-deterministic changes. Online Style Inference also generates more complete and longer formal translations. . . . .	130

## List of Figures

1.1	Formality Transfer (FT). It models the transformation from sentence $\mathbf{Y}_{\bar{\ell}}$ to sentence $\mathbf{Y}_{\ell}$ of the same language but at the opposite formality level $\ell$ . . . . .	4
1.2	Formality-Sensitive Machine Translation (FSMT). Given a sentence $\mathbf{X}$ and a desired formality level $\ell$ , it outputs a translation $\mathbf{Y}_{\ell}$ of the desired formality. . . . .	4
3.1	Train accuracy of formal/informal words classification and test accuracy of CTRW word-pair ranking vs. the (sub)space dimensionality. An SVM-based formality model achieved the best test performance on subspaces identified by PCA on PPDB data. . . . .	38
3.2	Test accuracy of CTRW word-pair ranking vs. the subspace dimensionality. All formality models achieved similar performance on subspaces of size 9-21 identified by PCA-PPDB. . . . .	42
5.1	The framework of bi-directional NMT with synthetic parallel data. A bi-directional model (Model-1) is initialized on parallel data, and it translates select source and target monolingual data. Training is then continued on the augmented parallel data, leading to a cycle of improvement ( $\rightarrow$ Model-2 $\rightarrow$ Model-3). . . . .	61
5.2	BLEU scores for four translation directions vs. the size of selected monolingual data. $n$ in x-axis equals to the size of real parallel data. EN $\rightarrow$ SW models use BLEU in parentheses in y-axis. Both language pairs tend to reach the plateau with more synthetic parallel data. . . . .	71
5.3	Training curves of perplexity on the training and the development sets for TR $\leftrightarrow$ EN. Reconstructing from hidden states (HIDDEN) and reconstructing from sampled translations ( $\beta = 0$ ) are compared. HIDDEN achieves extremely low training perplexity and suffers from unstable training during the early stage. . . . .	82

6.1	System overview: Our multi-task learning model can perform both bi-directional English formality transfer and translate French to English with desired formality. It is trained jointly on monolingual formality transfer data and bilingual translation data. . . . .	85
6.2	The training data used for multi-task learning models. The bi-directional formality transfer data and the bilingual data (e.g., FR-EN) of equivalent size are always concatenated. . . . .	88
6.3	BLEU improvements or scores for four transfer/translation directions vs. the size of FR-EN parallel data. $n$ in x-axis equals to the original size of bi-directional style transfer training data. Formality transfer improves with bilingual data and the performance reaches the plateau quickly. The translation quality increases monotonically with the size of training data. . . . .	98
7.1	Online Style Inference. Given a translation example $(\mathbf{X}, \mathbf{Y})$ , FSMT produces both informal and formal translations of $\mathbf{X}$ , i.e., $\mathbf{Y}_I = \text{FSMT}(\mathbf{X}, \ell_I)$ and $\mathbf{Y}_F = \text{FSMT}(\mathbf{X}, \ell_F)$ . $\mathbf{Y}$ is labeled as formal since it is closer to $\mathbf{Y}_F$ than $\mathbf{Y}_I$ . . . . .	112
7.2	Comparing $S_1$ and $S_2$ with LEPoD: hollow circles represent non-exact matched tokens, yielding a LED score of $(\frac{7}{15} + \frac{4}{12}) \times \frac{1}{2} = 0.4$ . Given the alignment illustrated above, the PoD score is $\frac{0+3+2+0}{10} = 0.5$ . . . . .	122
7.3	Win/Tie/Loss counts when comparing Online Style Inference to Multi-Task. Informal translations generated by OSI are annotated as more informal than Multi-Task, while formal translations are annotated as more formal. The OSI model also gets more instances that better preserve the meaning. . . . .	127

## Chapter 1: Introduction

Written and spoken language carry information beyond their literal meaning, such as the situation in which they might be used. For instance, while one can start a conversation with a friend on WhatsApp by saying “Hey Dude”, a formal letter is more likely to start with “Dear Sir or Madam”. Speakers’ choice of words and grammar conveys important information about the situational context and speaker purpose that listeners can interpret and respond to (Hovy, 1987; Biber, 1995). The resulting language variations are named *register variations*, or more broadly, *stylistic variations* — the latter also interprets linguistic differences that are not directly functional, such as dialect variations (Schilling-Estes, 2002; Biber and Conrad, 2009).

Computational models of natural language should consider both its meaning and style. We aim to control style in applications that generate language, with a focus on two common tasks in our daily life. The first task is text rewriting. Professional editors tailor or rewrite the text, and this procedure involves polishing and catering it to the target audience with proper stylistic features, besides correcting errors and improving readability. The second task is translation. Translations do not necessarily obey the conventions of the source language, such as register profiles

of the source (Lapshinova-Koltunski and Vela, 2015). Human translators translate a document for a specific audience (Nida and Taber, 2003), and often ask what the expected tone of the content is when taking a new translation job.<sup>1</sup> However, this type of style information is not taken into account in modern machine translation.

Among the many dimensions of stylistic variations, this dissertation focuses on textual formality. While textual style is also reflected along other dimensions of variations, including complexity or specificity, formality is considered a key dimension of style (Heylighen and Dewaele, 1999) and register variations (Biber, 2014), and it encompasses a range of finer-grained dimensions including politeness, seriousness and respect distinctions (Irvine, 1979; Brown and Fraser, 1979).

Incorporating stylistic aspects in natural language generation has been discussed for decades, but many early works proposed rule-based generation systems, which are not scalable (e.g., McDonald and Pustejovsky, 1985; Hovy, 1987; Power et al., 2003; Reiter and Williams, 2010; Mairesse and Walker, 2011). More recent work starts to leverage neural models, but style annotations are still acquired using rules (e.g., Fidler and Goldberg, 2017).

Automatic stylistic text rewriting (a.k.a. textual style transfer) is an emerging research area. A machine translation model is usually used if parallel texts with diverse styles are accessible for training (e.g., Xu et al., 2012; Zhang and Lapata, 2017; Rao and Tetreault, 2018; Carlson et al., 2018). Text rewriting models often fail by altering meaning in addition to style, especially when parallel texts are not

---

<sup>1</sup>A web-based human translation platform, Gengo, gives an example in the tutorial: <https://support.gengo.com/hc/en-us/articles/231438047>.

available. Research on unsupervised text rewriting is still in its infancy (e.g., [Mueller et al., 2017](#); [Hu et al., 2017](#); [Shen et al., 2017](#); [Fu et al., 2018](#)).

Controlling the style of machine translation (MT) output, which can be viewed as cross-lingual style transfer, is under-investigated. The pioneering work by [Di-Marco and Mah \(1994\)](#) and [Mima et al. \(1997\)](#) improves rule-based MT by analyzing syntactic stylistics or the speaker’s role and gender. In data-driven MT frameworks, style is not modeled explicitly. When a style is considered, it is equated with a domain or a provenance. For example, [Lewis et al. \(2015\)](#) and [van der Wees et al. \(2016\)](#) build conversational MT systems by selecting conversation-like training data; [Michel and Neubig \(2018\)](#) build personalized MT systems by using speaker-annotated TED talks. Prior work has also focused on narrow realizations of stylistic variations, such as T-V pronoun selection for translation into German ([Brown and Gilman, 1960](#); [Sennrich et al., 2016a](#)), or controlling the active/passive voice ([Yamagishi et al., 2016](#)).

## 1.1 Research Problems

This dissertation addresses formality style transfer within and across languages and shows that jointly modeling these two tasks helps address the limited availability of training data.

Formality style transfer **within** languages refers to the task of monolingual formality transfer (Figure 1.1), e.g., converting the informal sentence “What’s up?” to a formal one: “How are you doing?” It models the transformation from sentence



Figure 1.1: Formality Transfer (FT). It models the transformation from sentence  $\mathbf{Y}_{\bar{\ell}}$  to sentence  $\mathbf{Y}_{\ell}$  of the same language but at the opposite formality level  $\ell$ .



Figure 1.2: Formality-Sensitive Machine Translation (FSMT). Given a sentence  $\mathbf{X}$  and a desired formality level  $\ell$ , it outputs a translation  $\mathbf{Y}_{\ell}$  of the desired formality.

$\mathbf{Y}_{\bar{\ell}}$  to sentence  $\mathbf{Y}_{\ell}$  of the same language but at the opposite formality level  $\ell$ :

$$\hat{\mathbf{Y}} = \arg \max_{\mathbf{Y}_{\ell}} P(\mathbf{Y}_{\ell} | \mathbf{Y}_{\bar{\ell}}, \ell). \quad (1.1)$$

Formality transfer models are trained with monolingual sentence pairs that express the same meaning at different formality levels. These examples rarely occur naturally and are therefore only available in small quantities.

To study formality transfer **across** languages, we introduce a new task of controlling output formality in machine translation. For example, the French sentence “Comment ça va?” could be translated formally to “How are you doing?”, but we could also produce an informal equivalent “What’s up?” We define the task of *Formality-Sensitive Machine Translation* (FSMT, Figure 1.2), which takes two inputs, a sentence  $\mathbf{X}$  and a desired formality level  $\ell$ , and outputs a translation  $\mathbf{Y}_{\ell}$  of the desired formality. It can be modeled as follows:

$$\hat{\mathbf{Y}} = \arg \max_{\mathbf{Y}_{\ell}} P(\mathbf{Y}_{\ell} | \mathbf{X}, \ell). \quad (1.2)$$

The ideal training data for this task consists of translations of the same input in different styles, e.g.,  $(\mathbf{X}, \mathbf{Y}_{\ell_1}, \mathbf{Y}_{\ell_2})$ . Unfortunately, such data is not available.

We take a unified view of these two tasks: given a sentence expressed in English or a foreign language as the input, we generate an English sentence at the desired formality level automatically. The generated sentence should be fluent and preserve the meaning of the input. Controlling style requires being able to detect stylistic variations in text, such as annotating training examples for MT systems. The annotation could be either classification (e.g., informal vs. formal) or scoring (e.g., continuous formality level). Unlike politeness in German (i.e., T-V pronoun distinction) and active/passive voice in English, formality and many other styles cannot easily be labeled using rules.

We design systems to address this task based on the following hypotheses:

- Formality variations for language generation can be learned/modeled from examples, such as a pool of formal and informal words or sentence pairs.
- Joint modeling of formality transfer and machine translation improves formality transfer within and across languages, despite the limited nature and quantity of annotated style data. In particular, joint modeling enables FSMT without dedicated FSMT training examples.

## 1.2 Roadmap

This dissertation is organized into eight chapters. We discuss relevant background work on machine translation and style transfer in Chapter 2. Then, Chapters

3 to 7 describe our contributions. The final Chapter 8 concludes with a discussion of limitations and future work. We summarize each chapter below.

### 1.2.1 Modeling Lexical Stylistic Variations

We model lexical stylistic variations by placing words on a continuous formality scale. We hypothesize that differences between distributional representations of words that share the same meaning are indicative of style differences. To test this hypothesis, we identify salient dimensions of variations (i.e., a stylistic subspace) between word representations of lexical paraphrases. Evaluation on a formality prediction task demonstrates the benefits of using induced stylistic subspaces. We describe this method in Chapter 3.

### 1.2.2 Controlling Formality in Phrase-Based MT

Given formality annotations derived from modeling stylistic variations, we are now able to control the formality of machine translation output. We implement the initial FSMT system based on a standard phrase-based MT architecture. We first adapt our lexical style model to quantitatively measure formality levels of sentences. The resulting formality model provides the most accurate scores on intrinsic formality datasets. We then implement FSMT by  $n$ -best reranking. The reranking module promotes translation hypotheses whose formality levels are closer to the user-provided formality level (i.e., desired formality level).

Automatic and human evaluation suggest the effectiveness of our system in

controlling language formality without loss in translation quality. However, the space of possible outputs is limited to  $n$ -best translation hypotheses. We introduce the reranking-based FSMT in Chapter 4.

### 1.2.3 Low-Resource Neural Machine Translation

While lexical formality models estimate sentential formality by aggregating local information, neural models provide a more promising approach to model formality of sentences. *Neural Machine Translation* (NMT) has become the new standard of MT as it consistently outperforms previous methods across domains and language pairs (Bojar et al., 2017; Cettolo et al., 2017). The success of controlling politeness in NMT (Sennrich et al., 2016a) and using NMT for style transfer (Jhamtani et al., 2017; Zhang and Lapata, 2017; Rao and Tetreault, 2018) suggests that neural models are also well suited to our tasks.

Formality style transfer can be viewed as a low-resource MT problem given a limited number of parallel examples with diverse formality styles. We first research how to improve the translation quality of low-resource NMT independently of style by making better use of limited training data. (1) We first propose a bi-directional NMT framework inspired by multi-task learning. It trains both directions of a language pair jointly with a single model. Joint training can leverage limited training data effectively via duplication. The bi-directional model consistently achieves improved translation quality, particularly in low-resource scenarios. (2) We further introduce a differentiable input reconstruction loss to bi-directional NMT, aiming at

exploiting the source side of parallel samples. This loss compares original inputs to reconstructed inputs, which is obtained by back-translating translation hypotheses into the input language. This approach achieves small but consistent improvements on translating low-resource language pairs. Detailed description of the bi-directional NMT and the differentiable input reconstruction loss are presented in Chapter 5.

#### 1.2.4 Joint Model of Neural Formality Transfer and FSMT

We apply the bi-directional model from our low-resource NMT research to formality transfer tasks. Using the idea of bi-directional models yields an elegant and unified model that transfers between formal and informal language. The resulting models outperform uni-directional models, which matches the behavior of bi-directional NMT in low-resource settings.

We further adapt the idea of multi-task training to the FSMT task by jointly training bi-directional formality transfer and machine translation. The training shares information from two distinct types of supervision we can provide: sentence pairs in the same language that capture formality difference, and translation pairs drawn from corpora of diverse formality.

Experimental results show that the integrated neural model is able to perform FSMT without being explicitly trained on style-annotated translation examples. The joint model also achieves state-of-the-art performance for formality transfer. We present the neural formality transfer and FSMT via multi-task learning in Chapter 6.

### 1.2.5 Neural FSMT with Synthetic Supervision

Building an FSMT system ideally requires training triplets consisting of a bilingual sentence pair labeled with target language formality. The multi-task FSMT model, however, is presented with samples where one element of the triplet is always missing. Therefore, it sometimes produces translations without expected formality properties or formality-controlled outputs disobeying the source meaning.

We hypothesize that exposing the models to complete training triplets should further help formality-sensitive language generation: formal and informal outputs differ from each other and formality rewrites do not introduce translation errors. To this end, we introduce a new training scheme for multi-task FSMT models that automatically generates synthetic training triplets by inferring the target formality for a given parallel sentence pair during training.

Comprehensive automatic and human assessments show that our best model trained with synthetic supervision outperforms prior neural FSMT models. It produces translations that better match desired formality levels while preserving source meaning. We introduce our approaches to generate synthetic training triplets and analyze outputs qualitatively to illustrate how formality is marked in model outputs in Chapter 7.

### 1.3 Contributions

This dissertation makes the following contributions:

- We model lexical formality by learning style dimensions in word embedding spaces based on variations between embeddings of paraphrases. The induced style subspace better distinguishes more formal from less formal words than the original space (Niu and Carpuat, 2017).
- We introduce a new task, Formality-Sensitive Machine Translation, and design a statistical and reranking-based system to perform French-English FSMT using lexical formality scores (Niu et al., 2017).
- We design neural systems using multi-task learning that address formality transfer and FSMT jointly. They achieve state-of-the-art performance on English formality transfer and perform French to English FSMT without being explicitly trained on style-annotated translation examples (Niu et al., 2018b).
- We further improve the zero-shot multi-task learning approach with synthetic supervision. After being trained with complete training triplets, this FSMT system produces translations that better match desired formality levels while preserving the source meaning (Niu and Carpuat, 2019).
- We improve low-resource neural machine translation by introducing (1) a bi-directional model which performs iterative back-translation without auxiliary models (Niu et al., 2018a), and (2) a differentiable input reconstruction loss

which exploits the source side of parallel samples without additional parameters (Niu et al., 2019).

- We release training scripts for aforementioned systems and implementations of our new training objectives at <https://github.com/xingniu>.

## Chapter 2: Background

This work focuses on machine translation and style transfer tasks that can both be framed as sequence-to-sequence transformations. Machine translation is the fundamental framework we build on and we review related concepts and techniques in Section 2.1. Controlling style in machine translation output or any other language generation task requires modeling stylistic variations. We introduce related work in modeling styles with a focus on formality in Section 2.2 and review how stylistic variations are incorporated with language generation models in Section 2.3.

### 2.1 Machine Translation

*Machine Translation* (MT) is the task of using computers to translate from one natural language into another. Data-driven approaches to MT have dominated both research and commercial market by learning translation patterns from large parallel corpora, which are bilingual corpora containing original documents and their translations produced by humans. *Statistical Machine Translation* (SMT) provided the first family of architectures (Brown et al., 1993; Berger et al., 1994; Lopez, 2008; Koehn, 2010), while neural models have recently gained traction (Bahdanau et al., 2015; Wu et al., 2016; Hassan et al., 2018).

Mathematically, we formulate the translation probability for translating a sentence of the source language  $\mathbf{X} = (x_1, \dots, x_n)$  into the target language  $\mathbf{Y} = (y_1, \dots, y_m)$  as  $P(\mathbf{Y}|\mathbf{X})$ . Given an MT model with parameters  $\theta$ , and an input sentence  $\mathbf{X}$ , the MT task consists in finding the most probable translation, i.e.

$$\hat{\mathbf{Y}} = \arg \max_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{X}; \theta). \quad (2.1)$$

Among a taxonomy of various MT approaches, phrase-based models and neural models draw most attention. We implement cross-lingual and monolingual formality transfer systems based on these two models.

### 2.1.1 Phrase-Based Machine Translation

*Phrase-Based Machine Translation* (PBMT) models translations at the granularity of contiguous sequences of words, called phrases, between source and target languages using statistical methods (Och et al., 1999; Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004). PBMT is usually formulated as a log-linear model (Och and Ney, 2002),

$$P(\mathbf{Y}|\mathbf{X}; \theta) = \frac{\exp \sum_k \lambda_k h_k(\mathbf{X}, \mathbf{Y})}{\sum_{\mathbf{Y}'} \exp \sum_k \lambda_k h_k(\mathbf{X}, \mathbf{Y}')}, \quad (2.2)$$

where  $h_k(\mathbf{X}, \mathbf{Y})$  are feature functions while  $\lambda_k$  are feature weights. Some core features are defined over decompositions of sentences  $\mathbf{X}$  and  $\mathbf{Y}$  into phrases, which are learned from word-level alignments (Brown et al., 1993).

Training PBMT models is a combination of generating features on the training set and estimating feature weights on the tuning sets. Generating features

involves training multiple preliminary models independently. For example, phrase translation probabilities are aggregated from word-level alignments, which are also automatically learned from bitext (Marcu and Wong, 2002; Och and Ney, 2003); target language models are built from monolingual corpora to encourage generating fluent output. Other crucial features include word reordering that captures language differences in word order (Koehn et al., 2005; Galley and Manning, 2008) and word penalty that calibrates the output length (Koehn et al., 2003), etc. Estimating feature weights in Equation 2.2 is intractable, because computing the denominator involves getting all possible translations. Therefore, this sum is usually approximated over the  $n$ -best output (Och and Ney, 2002). Practically, PBMT models are trained by maximizing a translation quality measurement, e.g., BiLingual Evaluation Understudy (BLEU, Papineni et al., 2002), by using optimization algorithms such as *Minimum Error Rate Training* (MERT, Och, 2003), *Margin Infused Relaxed Algorithm* (MIRA, Crammer and Singer, 2003) and batch MIRA (Cherry and Foster, 2012).

Generating translations, called decoding, is a search procedure that aims to find a sequence of phrases with maximum probability estimated by PBMT models. Collecting all combinations of phrases is intractable, so beam search is usually employed as an approximation and helps balance efficiency with exploring multiple translation options beyond greedy search (Koehn et al., 2003; Koehn, 2004a).

In order to incorporate features that target specific model errors but may not be efficiently computed in the decoder, Och et al. (2004) and Shen et al. (2004) propose to rerank  $n$ -best translation hypotheses by inputting them to an auxiliary

model with access to additional feature functions, such as alternative alignment scores, language models and rules. We leverage this technique to re-select translations matching expected formality levels.

## 2.1.2 Neural Machine Translation

*Neural Machine Translation* (NMT) parameterizes the probability  $P(\mathbf{Y}|\mathbf{X})$  as a single large neural network with parameters  $\theta$ , that can be trained end-to-end. NMT can be viewed as a conditional language model, where the probability of the target word  $y_t$  at step  $t$  is conditioned on the target history  $\mathbf{Y}_{<t} = (y_1, \dots, y_{t-1})$  and the source sentence  $\mathbf{X}$ . So the probability of the target sequence in Equation 2.1 is factorized as

$$P(\mathbf{Y}|\mathbf{X}; \theta) = \prod_t P(y_t|\mathbf{Y}_{<t}, \mathbf{X}; \theta). \quad (2.3)$$

The right-hand side probability of Equation 2.3 is parameterized via an encoder-decoder neural network (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014). Words in a sentence are first mapped to vector representations (a.k.a. embeddings). We reuse  $\mathbf{X}$  or  $\mathbf{Y}$  to represent a sentence as a sequence of embeddings for simplicity’s sake. The encoder transforms a source sentence to a sequence of hidden states  $\mathbf{S} = (\mathbf{s}_1, \dots, \mathbf{s}_n)$ :

$$\mathbf{S} = \text{Encoder}(\mathbf{X}). \quad (2.4)$$

The decoder produces an hidden state  $\mathbf{h}_t$  at each step  $t$ , given previous hidden state  $\mathbf{h}_{t-1}$  and/or target word embeddings  $\mathbf{Y}_{<t}$ , and a context vector  $\mathbf{c}_t$ :

$$\mathbf{h}_t = \text{Decoder}(\mathbf{h}_{t-1}, \mathbf{Y}_{<t}, \mathbf{c}_t). \quad (2.5)$$

The encoder and decoder can be implemented by various neural network architectures, such as *Recurrent Neural Networks* (RNNs), to handle variable-length sequences. [Cho et al. \(2014\)](#) and [Sutskever et al. \(2014\)](#) use Gated Recurrent Units (GRUs) and Long Short-Term Memory cells (LSTMs, [Hochreiter and Schmidhuber, 1997](#)) respectively to realize RNNs.

The context vector  $\mathbf{c}_t$  is calculated via an attention mechanism ([Bahdanau et al., 2015](#); [Luong et al., 2015](#)) by querying an intermediate hidden state  $\tilde{\mathbf{h}}_t$  to source hidden states  $\mathbf{S}$  and computing a weighted sum of source hidden states:

$$\mathbf{c}_t = \text{Attention}(\mathbf{h}_t, \mathbf{S}) \tag{2.6}$$

$$= \text{softmax}(\alpha(\tilde{\mathbf{h}}_t, \mathbf{S})) \cdot \mathbf{S}, \tag{2.7}$$

where  $\alpha$  produces a similarity matrix, such as using dot product.

Finally, the probability per token is estimated by a softmax output layer over a linear transformation that transforms  $\mathbf{h}_t$  to a distribution over the vocabulary:

$$P(\cdot | \mathbf{Y}_{<t}, \mathbf{X}; \boldsymbol{\theta}) = \text{softmax}(\mathbf{W}\mathbf{h}_t + \mathbf{b}), \tag{2.8}$$

where  $\mathbf{W}$  and  $\mathbf{b}$  are the weight matrix and bias vector respectively.

RNN-based sequence to sequence models have some disadvantages. On one hand, contextual information fades along the long sequential process. On the other hand, the representation at each time step is dependent upon its precursor, which limits parallelization. Recently, [Gehring et al. \(2017\)](#) replace RNNs with *Convolutional Neural Networks* (CNNs): a convolutional layer combines the context in a limited window into a single representation. The effective window size grows when

stacking multiple layers. Vaswani et al. (2017) model all dependencies by the attention mechanism, which is time-independent. Besides attentions between decoder states and encoder states, they propose *self-attention* that computes the association between any word and any other word (or any previously produced word for decoder) in the same sequence.

Regardless of the specific architecture chosen, the standard training objective for NMT is to maximize the *log-likelihood* of the training data:

$$\mathcal{L}_{MT} = \sum_{(\mathbf{X}, \mathbf{Y})} \log P(\mathbf{Y} | \mathbf{X}; \boldsymbol{\theta}) \quad (2.9)$$

$$= \sum_{(\mathbf{X}, \mathbf{Y})} \sum_t \log P(y_t | \mathbf{Y}_{<t}, \mathbf{X}; \boldsymbol{\theta}). \quad (2.10)$$

Maximizing  $\mathcal{L}_{MT}$  is equivalent to minimizing the cross-entropy between the predicted softmax distribution (Equation 2.8) for  $\hat{y}_t$  and the ground truth one-hot distribution for  $y_t$ . Popular optimization algorithms for NMT are Stochastic Gradient Descent (SGD, Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952) and Adam (Kingma and Ba, 2015).

Calculating the conditional probability in Equation 2.10 during training is realized by using the *teacher forcing* strategy (Williams and Zipser, 1989), which always feeds in the ground truth previous tokens  $\mathbf{Y}_{<t}$  when predicting the current token. During the test time, the model relies on its own predictions to generate translations. It aims to find a sequence of tokens with maximum probability and also uses beam search as an approximation, which is the same as PBMT.

### 2.1.3 Multilingual and Zero-Shot Neural Machine Translation

The designs of our cross-lingual and monolingual formality transfer systems are inspired by *Multi-Task Learning* (MTL), which is used for transferring domain knowledge between related tasks (Caruana, 1997). MTL has been found to be beneficial for several natural language processing tasks in past work, ranging from part-of-speech tagging and parsing to query classification and document ranking (Collobert and Weston, 2008; Liu et al., 2015; Luong et al., 2016). Approaches based on neural networks can leverage cross-task data (e.g., datasets of multiple sequence tagging tasks) by learning shared representations or layers to improve generalization.

Multi-task learning has been successfully used to build multilingual translation models, in which parallel training corpora of various language pairs are concatenated and certain components are shared. A one-to-many translation system can be built by sharing both the encoder and the attention mechanism (Dong et al., 2015). A many-to-many translation system can be built by sharing only the attention (Firat et al., 2016). Surprisingly, Johnson et al. (2017) enable a standard NMT framework to support many-to-many translation directions by simply attaching a special token (indicating the target language) to each source sentence. They also report promising results for translation between languages that have zero parallel data (a.k.a. zero-shot translation). We investigate a special case of multilingual translation, bi-directional translation and transfer, in low-resource settings.

We build formality transfer and machine translation jointly to perform zero-

shot FSMT without training on bilingual parallel data with formality annotations. The resulting model is similar to zero-shot multilingual NMT and they both face the challenge that, for example, the zero-shot translation usually performs worse than supervised models and even the simple pivoting approach which leverages an intermediary language as the bridge (Johnson et al., 2017). There have been efforts to improve this vanilla strategy by filtering out vocabulary entries of incorrect languages prior to translation (Ha et al., 2017), using a dedicated attention module per target language (Blackwood et al., 2018), contextually generating dedicated encoder-decoder parameters for any language pair (Platanios et al., 2018), using an auxiliary loss to encourage encoding sentences into source-language invariant representations (Arivazhagan et al., 2019), and encouraging the model to produce equivalent translations of parallel sentences into an auxiliary language (Al-Shedivat and Parikh, 2019). From a different angle, we tackle this problem by automatically inferring labels.

#### 2.1.4 Round-Trip Neural Machine Translation

Optimizing NMT models by maximizing the log-likelihood (Equation 2.9) works well when abundant training data is available, but it is still an open question how to best train deep neural models from limited parallel data. As alternatives to combining parallel data of multiple language pairs with standard training, we will discuss training strategies inspired by the idea of round-trip translation: suppose input sentence  $\mathbf{X}$  is translated forward to  $\hat{\mathbf{Y}}$  and then translated back to  $\hat{\mathbf{X}}$ , then

$\hat{\mathbf{Y}}$  is more likely to be a good translation if the distance between  $\hat{\mathbf{X}}$  and  $\mathbf{X}$  is small (Brislin, 1970).

Using the round-trip translation (also called input reconstruction) as a training signal for NMT usually requires a complex training process with reinforcement learning and auxiliary models to perform back-translation. For instance, Cheng et al. (2016) add a reconstruction loss for monolingual examples to the training objective. He et al. (2016) in addition evaluate the quality of  $\hat{\mathbf{Y}}$  by a language model. Both approaches have symmetric forward and backward translation models which are updated alternatively. This requires policy gradient algorithms for training, which are not always stable.

Back-translation (Sennrich et al., 2016b) is a simpler yet effective strategy, which performs only half of the reconstruction process: it generates a synthetic source side for monolingual target language examples  $\mathbf{Y} \rightarrow \hat{\mathbf{X}}$ . It uses an auxiliary backward model to generate the synthetic data but only updates the parameters of the primary forward model. Wang et al. (2018c) extend this method by generating  $K$  synthetic source sentences and minimizing the difference between  $P(\mathbf{Y})$  and  $\mathbb{E}_{\hat{\mathbf{X}}}P(\mathbf{Y}|\hat{\mathbf{X}};\theta)$ . Iteratively updating forward and backward models (Zhang et al., 2018; Hoang et al., 2018; Cotterell and Kreutzer, 2018) is an expensive solution as back-translations are regenerated at each iteration.

Aforementioned methods produce intermediate (synthetic) translations using beam search, but beam search is not differentiable which prevents back-propagating reconstruction errors. Prior work has sought to simplify the optimization of reconstruction losses by side-stepping beam search. Tu et al. (2017) first propose to re-

construct NMT input from the decoder’s hidden states while [Wang et al. \(2018a,b\)](#) suggest to use both encoder and decoder hidden states to improve translation of dropped pronouns. However, these models may achieve low reconstruction errors by learning to copy the input to hidden states. To avoid copying the input, [Artetxe et al. \(2018\)](#) and [Lample et al. \(2018a\)](#) use denoising autoencoders ([Vincent et al., 2008](#)) in unsupervised NMT. We will introduce a simple and effective alternative in [Chapter 5](#).

## 2.2 Stylistic Variations

Stylistic variations reflect differences in language (such as changes of vocabulary and syntactic structures) associated with situational contexts or purposes. We first discuss formality, the prime dimension of stylistic variation we investigate. We then review how stylistic variations are computationally modeled.

### 2.2.1 Formality: Definition and Discussion

People can make intuitive distinction between formal language (e.g., an essay) and informal language (e.g., an instant message) without referring to a conceptual definition of “formality”. Based on extrinsic characteristics, [Richards et al. \(1997\)](#) define “formal speech” in a dictionary as “the type of speech used in situations when the speaker is very careful about pronunciation and choice of words and sentence structure.” [Heylighen and Dewaele \(1999\)](#) hypothesize that people invest more than the normal attention in the form of expressions because they want to make sure

that their expressions are not misunderstood, and define “formal” as “attention to form for the sake of unequivocal understanding of the precise meaning of the expression.” However, neither of these definitions can be interpreted into comprehensive guidelines without introducing subjective opinions, let alone used for computational assessments. In the present dissertation, we define “formality” by aggregating a significant amount of examples, which are collected from human-annotated datasets or corpora.

Recent research shows factors affecting formality via crowdsourcing. [Pavlick and Tetreault \(2016\)](#) and [Rao and Tetreault \(2018\)](#) ask annotators to rewrite informal sentences (from Yahoo Answers) in order to make them more formal. Common types of edits made in rewriting include capitalization, punctuation, phrasal paraphrasing, deletion of fillers, completion, expansion of contractions, spelling correction, normalization, etc.

People use various formality levels when addressing different audiences because using the formal language is not always superior. Despite having less chance to be misinterpreted by others who do not share the same context as the sender, formal speech bears the disadvantages of being more static or rigid, and structurally complex.

## 2.2.2 Modeling Stylistic Variations

Modeling stylistic variations is important for building style-aware systems. By scoring text in terms of styles, we are able to annotate training data and perform

automatic evaluations.

Many studies of stylistic variations have focused on the corpus or sentence level. For instance, multidimensional corpus analysis (Biber, 1995) relies on statistical analysis to identify the salient linguistic co-occurrence patterns that underlie register variations. Heylighen and Dewaele (1999, 2002) define the characteristics of formality and quantitatively represent formality levels by word frequencies per part-of-speech tags. More recently, richer combinations of features have been used to measure formality. Li et al. (2013) leverage lexicons, part-of-speech classifiers, syntactic parsers, templates, etc. to capture formality features such as narrativity, cohesion, syntactic simplicity and word correctness. Pavlick and Tetreault (2016) provide a thorough study of sentence-level formality and show that classifiers based on features including part-of-speech tags and dependency parses can predict formality as defined by the collective intuition of human annotators.

We focus on identifying dimensions of lexical stylistic variations. Prior work on evaluation of style factors at the word level has used standard word embeddings as features, and relied on external supervised methods to identify style relevant information in these embeddings. Brooke et al. (2010) propose to score the formality of a word by comparing its meaning to that of seed words of known formality using cosine similarity (Turney and Littman, 2003). Rothe and Schütze (2016) and Rothe et al. (2016) show that meaningful ultradense subspaces that capture dimensions such as polarity and concreteness can be induced from word embeddings in a supervised fashion.

Other approaches include work by Pavlick and Nenkova (2015) who uses a un-

igram language model to capture the difference between lexical distributions across genres. [Preotiuc-Pietro et al. \(2016\)](#) isolate stylistic differences associated with user attributes (e.g., gender and age) by using paraphrase pairs and word distributions similar to [Pavlick and Nenkova \(2015\)](#). Analysis of stylistic variations from the point of view of the lexicon also includes predicting term complexity, as annotated by non-native speakers ([Paetzold and Specia, 2016](#)).

## 2.3 Stylistic Variations in Language Generation

Our ultimate goal is generating language with specified target style (formality). We first review techniques used for related tasks of language generation conditioned on certain properties or semantics of interests. Then, we introduce related work on how to model stylistic variations jointly with our focused task, which is text generation within and across languages. Finally, we briefly overview the evaluation methodology for these tasks.

### 2.3.1 Conditional Language Generation

Language generation can be controlled by various aspects, including styles. Most recently, RNN models demonstrate their potential. In order to generate dialogues conditioned on semantic information, [Wen et al. \(2015\)](#) introduce a control cell into LSTM to gate the dialogue act. [Li et al. \(2016\)](#) propose persona-based neural conversation models in which speaker (or speaker-addressee) embeddings are learned jointly with word embeddings and generated conversation responses are

conditioned on the speakers’ identities. Similarly, [Ficler and Goldberg \(2017\)](#) use a conditional neural language model to control linguistic style aspects in language generation, by appending a pre-defined style vector to each predicted word embedding vector in the target sequence. [Kikuchi et al. \(2016\)](#) focus on controlling the output length and propose either appending the remaining length to the LSTM input at each step, or multiplying the desired length to initial LSTM cell state. These approaches point out possible strategies (i.e., manipulating decoders’ RNN states or embeddings) in injecting formality information when generating languages.

### 2.3.2 Textual Style Transfer or Rewriting

Monolingual textual style transfer or rewriting is a sub-area of conditional language generation — new text is generated via paraphrasing while conditioned on style changing. Style transfer includes two essentials: (1) The input and stylized output must share identical semantic (non-stylistic) content; (2) The transformation must produce desired stylistic shifts.

Style transfer can naturally be framed as a sequence to sequence translation problem given sentence pairs that are paraphrases in two distinct styles. These parallel style corpora are constructed by creatively collecting existing texts of varying styles, and are therefore rare and much smaller than machine translation parallel corpora. For instance, [Xu et al. \(2012\)](#) scrape modern translations of Shakespeare’s plays and use a PBMT system to paraphrase Shakespearean English into/from modern English. [Jhamtani et al. \(2017\)](#) improve performance on this dataset using NMT

with pointers to enable copy actions. The availability of parallel standard and simple Wikipedia (and sometimes additional human rewrites) makes text simplification a popular style transfer task, typically addressed using MT models ranging from syntax-based MT (Zhu et al., 2010; Xu et al., 2016), phrase-based MT (Coster and Kauchak, 2011; Wubben et al., 2012) to neural MT (Wang et al., 2016) trained via reinforcement learning (Zhang and Lapata, 2017). Another relatively easy-to-collect parallel style corpus is the Bible with various versions. Carlson et al. (2018) treat paraphrasing between Bible versions as monolingual translation and use the approach of multilingual NMT with side constraints (Johnson et al., 2017) to perform zero-shot bible style transfer.

Naturally occurring examples of parallel formal-informal sentences are harder to find. Prior work relied on synthetic examples generated based on lists of words of known formality (Sheikha and Inkpen, 2011). This state of affairs recently changed, with the introduction of the first large scale parallel corpus for formality transfer, Grammarly’s Yahoo Answers Formality Corpus (GYAFC, Rao and Tetreault, 2018). They collected over one hundred thousand informal sentences from Yahoo Answers and their formal rewrites via crowd-sourcing. They also presented benchmark style transfer systems based on both PBMT and NMT models. We leverage this corpus to enable multi-task monolingual and cross-lingual style transfer.

Another thread of research dealing with the lack of training data is unsupervised style transfer and it first achieves promising progress in computer vision. Gatys et al. (2016b,a) make pioneering work on migrating the semantic content of one image to different styles. They use CNNs to obtain both the source content

representation and the target style representation independently, and then generate style-transferred images by matching both representations. Neural style transfer for text is naturally more challenging than images. Unlike image pixels, words are discrete — a subtle shift in the continuous vector space could lead to another word being selected and might result in an unstable change in style and meaning.

Exploratory approaches for unsupervised textual style transfer or rewriting<sup>1</sup> have been proposed recently and many of them are based on autoencoders. [Mueller et al. \(2017\)](#) use a *Variational Auto-Encoder* (VAE, [Kingma and Welling, 2014](#)) to encode a sequence to a latent representation  $\mathbf{z}$ . They optimize  $\mathbf{z}$  until an expected stylistic score can be inferred from it. The transferred sequence is obtained by decoding optimized  $\mathbf{z}$ . [Hu et al. \(2017\)](#) use the same topology yet with several differences. First, they separate the latent representations into sequence embedding  $\mathbf{z}$  and stylistic code  $\mathbf{c}$ . Second, a discriminator is optimized to infer  $\mathbf{c}$  from a sequence instead of  $\mathbf{z}$ . [Shen et al. \(2017\)](#) use a VAE to encode sequences with different style labels into a shared latent space. Discriminators for different styles are optimized to distinguish real and transferred sequence. [Fu et al. \(2018\)](#) train two adversarial networks to enforce the meaning representations to be independent of style. The meaning representation is decoded in two ways: using multiple decoders or adding style-embeddings similar to [Hu et al. \(2017\)](#). The NMT framework is also borrowed by unsupervised style transfer. [Prabhumoye et al. \(2018\)](#) reduce stylistic properties of a sentence by translating it into another language. [Lample et al. \(2019\)](#) argue that

---

<sup>1</sup>Text rewriting includes tasks that change properties coupled with the meaning, such as sentiment transfer.

disentangling style from meaning is not necessary nor easily achievable in practice. They leverage back-translation to construct synthetic transfer pairs and use style embeddings as the initial state of the decoder for conditional language generation. We do not perform unsupervised textual style transfer in the present dissertation, but these explorations share inspiring ideas such as the importance of reconstruction accuracy in meaning preservation.

### 2.3.3 Controlling the Output Style in Machine Translation

Controlling the output style in machine translation has received sparse attention. The pioneering work by [DiMarco and Mah \(1994\)](#) and [Mima et al. \(1997\)](#) improves rule-based MT using linguistic features or extra-linguistic information such as speaker’s role and gender.

With the success of data-driven MT frameworks, people usually define styles by leveraging representative sub-data. For example, after selecting or annotating data of interest beforehand, [Lewis et al. \(2015\)](#) and [van der Wees et al. \(2016\)](#) build conversational MT systems, [Rabinovich et al. \(2017\)](#) build gender-specific MT systems. One of our baseline FSMT system is built with data selection.

Multiple-style annotated data further facilitates building a single NMT systems supporting translations of various styles. For example, [Michel and Neubig \(2018\)](#) build personalized NMT systems that optimize translation accuracy per each speaker. They achieve this by assigning a dedicated bias for each speaker trait in the output layer. [Korotkova et al. \(2018\)](#) train an NMT system with multiple sources of

data together and distinguish data sources by concatenating style vectors to source word embeddings (Sennrich and Haddow, 2016). The resulting system is able to translate the same input to various styles represented by the data source.

Sennrich et al. (2016a) show the first effort in controlling opposite styles for NMT. Specifically, they append a side constraint,  $\langle T \rangle$  or  $\langle V \rangle$  (i.e., T-V pronoun distinction), to the source text to indicate which pronoun is preferred in the German output (e.g., translating to polite *Sie* instead of informal *du/ihr* from the English word *you*). The T-V pronoun distinction only reflects one narrow dimension of the formality variation. Yamagishi et al. (2016) use the same method to control the active/passive voice of the translation. We employ this simple yet effective strategy to build our style-constrained neural language generation models. But the difference is that formality is difficult to be unambiguously annotated by artificial rules. Effectively modeling the formality variation for language generation is the challenge we face with.

### 2.3.4 Evaluation of Style-Constrained Language Generation

We evaluate both formality transfer and machine translation models by comparing the output against the human reference rewrites or translations and using BiLingual Evaluation Understudy (BLEU, Papineni et al., 2002). It is a precision-oriented metric in that it measures how much of the system output is correct, in terms of exact matches of  $n$ -grams. Formally, the most used BLEU with up to

4-gram matching is defined as

$$\text{BLEU} = \text{BP} \cdot \exp \left( \sum_{n=1}^4 \frac{1}{4} \log p_n \right), \quad (2.11)$$

where  $p_n$  is the geometric mean of the test corpus’  $n$ -grams precision,<sup>2</sup> and BP is the brevity penalty that penalizes scores if the system output is shorter than the references. BLEU is the *de facto* standard for automatic MT evaluation since it is easy to use and correlates highly with human evaluation.

However, BLEU is not an ideal automatic metric for FSMT because the reference translation given both an input sentence and a formality level is not available. Using translations with arbitrary style in standard MT test sets, BLEU may conflate mismatches due to translation errors and due to correct stylistic rewrites.

Despite being expensive, human evaluation is more reliable for language generation tasks (Hashimoto et al., 2019). Transfer intensity, content preservation and naturalness (fluency) are three key dimensions measured in text rewriting tasks (Mir et al., 2019). We follow this convention to evaluate both formality transfer and FSMT models.

---

<sup>2</sup>Precisely, it is a modified precision. Please refer to the original paper for details.

## Chapter 3: Discovering Lexical Stylistic Variations in Distributional Vector Space Models

Controlling style requires being able to detect stylistic variations in text, such as annotating training examples for MT systems. The annotation could be either classification (e.g., informal vs. formal) or scoring (e.g., continuous formality level). Unlike politeness in German (i.e., T-V pronoun distinction, [Brown and Gilman, 1960](#)) and active/passive voice in English, formality and many other styles cannot easily be labeled using rules.

In this chapter, we first investigate how stylistic variations are embedded in the topology of distributional vector space models and then use the produced style dimensions to place words on a continuous formality scale. Words are represented as dense vectors (i.e., word embeddings), and they have been showed to capture semantic similarity and other lexical semantic relations ([Mikolov et al., 2013](#); [Baroni et al., 2014](#); [Levy and Goldberg, 2014](#)).

We hypothesize that differences between embeddings of words that share the same meaning are indicative of style differences. For example, “watch” and “observe” are synonyms, but the latter is more formal. In order to test this hypothesis, we introduce a method based on *Principal Component Analysis* (PCA, [Pearson,](#)

1901; Hotelling, 1933) to identify salient dimensions of variations between word embeddings of lexical paraphrases. Applying this method to word embeddings learned from two large corpora representing distinct genres, we conduct a qualitative analysis of the principal components discovered. It suggests that the principal components indeed discover variations that are relevant to style.

Next, we evaluate the produced style dimensions (i.e., principal components) more directly, using them to distinguish more formal from less formal words. The formality prediction task lets us evaluate empirically the impact of different factors in identifying style-relevant dimensions, including dimensionality of the subspace and the nature of the prediction method. We also conduct an error analysis revealing the limitation of predicting formality based on vector space models.<sup>1</sup>

### 3.1 Approach

Our approach to discovering stylistic variations in vector space models is based on the assumption that these variations cannot be explained by differences in meaning, and they can be captured by salient dimensions of variation in the distributional spaces.

Lexical paraphrases should have the same meaning, and therefore their embeddings should be close to each other. When lexical paraphrases are not in the same location in the vector space, distances between them might be indicative of latent style variations. We discover such latent directions using PCA.<sup>2</sup>

---

<sup>1</sup>Code is available at <https://github.com/xingniu/computational-stylistic-variations>.

<sup>2</sup>Other algorithms for dimensionality reduction could also be used to discover latent variations, e.g., multidimensional scaling (MDS) and t-distributed stochastic neighbor embedding (t-SNE).

Concretely, suppose  $\mathbf{e}_i$  is the word embedding in the vector space for word  $w_i$ . Given pairs of word embeddings  $(\mathbf{e}_1, \mathbf{e}_2)$  for lexical paraphrases  $(w_1, w_2)$ , we subtracted them to get the relative direction  $\mathbf{d} = \mathbf{e}_1 - \mathbf{e}_2$ .

For a given word pair, the difference vector might capture many things besides style variations. We hypothesize that the regularities among these differences for a large number of examples will reveal stylistic variations. Therefore, we then trained a PCA model on all directional vectors to get principal components  $(\mathbf{pc}_k)$  capturing latent variations.

## 3.2 Qualitative Analysis of Latent Style Dimensions

### 3.2.1 Models Settings

The approach outlined above requires two types of inputs: (1) a word embedding space, and (2) a set of lexical paraphrases.

**Word Embeddings** We use word2vec (Mikolov et al., 2013) to build 300-dimensional vector space models for two corpora representing different genres. As suggested by Brooke et al. (2010), we select the ICWSM 2009 Spinn3r dataset (English tier-1) as the training corpus (Burton et al., 2009). It consists of about 1.6 billion words in 7.5 million English blogs and is expected to have wide variety of language genres. We also compare it with the pre-trained 300-dimensional model of Google News,<sup>3</sup> which represents an even larger training corpus but in a narrower register. By working

---

<sup>3</sup><https://code.google.com/archive/p/word2vec/>

with two different corpora, we aim to discover whether they share some common stylistic variations even though they have distinct word distributions.

**Lexical Paraphrases** PPDB 2.0 (Pavlick et al., 2015) provides automatically extracted lexical paraphrases with entailment annotations. We use the S-size pack and extract word pairs with **Equivalence** entailment relation, which represent a cleaner subset of the original PPDB. This process yields 9,427 paraphrase pairs found in the vocabulary of the blogs embeddings and 6,988 pairs found in the vocabulary of the Google news embeddings.

### 3.2.2 Analysis

We illustrate the principal components discovered in Table 3.1. For each of the top principal components, we can identify the most representative word pairs for that component by projecting all word pairs on  $\mathbf{pc}_k$  and ranking pairs based on  $\mathbf{d} \cdot \mathbf{pc}_k$ .

The first observation is that the first principal components for both blogs and news corpora capture the pattern of American/British-English variations (grey-boxed in the Table). These might also be related to the formality dimension of style, as British-English can be regarded to be more formal than American-English (Hurtig, 2006). However, not all representative word pairs fall in that category, and the nature of the variation between e.g., “annulling” and “canceling” is harder to characterize.

We can observe clues of stylistic variations in the subsequent (second+) prin-

$k$	Representative word pairs
	<b>ICWSM 2009 Spinn3r Blogs</b>
1	annulling • canceling    abolished • canceled    centre • center    emphasise • highlight programme • program    imperatives • essentials    motorway • freeway    labour • labor organised • organize    six-party • six-way    tranquility • serenity    tripartite • three-way
2	spendings • expenditures    summons • subpoenas    anti-malaria • antimalarial doctor • physician    falls • decreases    banned • prohibiting    fallen • decreased
3	decreased • receded    decreased • fallen    decreased • declined    decreased • shrank
4	agreements • understandings    unlimited • unbounded    disruptions • perturbations discriminatory • discriminative    timetable • time-scale    amended • altered    ban • forbidden
5	underscored • underline    eliminated • delete    highlights • underline    widened • expand widened • broaden    emphasises • underline    decreased • reduce    performed • fulfil
6	co-operate • collaborating    interdomain • cross-domain    cooperate • collaborating origin • sourcing    executions • implementations    multifunctional • cross-functional
7	refusing • rebuffs    stopped • halts    stress • underlines    inspected • reviewed withdrawals • withdraws    supervising • oversees    stress • emphasises    refused • rejects
8	restarting • revitalising    co-operation • collaborations    cooperation • collaborations restart • resumes    cleric • clergymen    cooperates • collaborates    expel • expulsions
9	obtain • gain    multi-factor • multifactorial    restricts • hampers    retrieves • recovers obstructs • hampers    revoking • canceling    contravened • breaches    invalidated • canceled
10	delete • eliminate    underline • stresses    underline • emphasises    schema • schemes restarting • revitalising    decreased • reduce    underline • highlight    permissions • permits
	<b>Google News</b>
1	educator • educationist    ousts • deposes    exemptions • derogations    educator • educationalist legal • juridical    truck • lorry    exceptions • derogations    accomplishments • attainments roadway • carriageway    prohibit • proscribe    freeway • motorway    lucrative • remunerative
2	standardize • standardizing    intercept • intercepting    evacuate • evacuating    isolate • isolating
3	destroys • demolishing    solves • resolving    impedes • obstructing    examines • investigating
4	falls • decreases    widens • increases    spends • expenditures    shrinks • decreases
5	infeasible • impracticable    impossible • impracticable    earmarks • allocates unworkable • impracticable    confines • restricts    impractical • impracticable

Table 3.1: Representative word pairs for top principal components (indexed by  $k$ ) are listed for both blogs and news corpora. A mixed variation of formality and American-British English (grey-boxed) can be characterized by the first principal component, but the following principal components seem vaguer in terms of interpreting stylistic variations.

principal components, but in general it is difficult to interpret each group. Several word pairs illustrate formality variations (e.g., “falls”  $\leftrightarrow$  “decrease”, “delete”  $\leftrightarrow$  “eliminate”). Many word pairs are literally exchangeable, but one in the pair is preferred under a specific context, such as “summons” vs. “subpoenas”, “decreased” vs. “fallen”, etc. Some principal components simply capture groups of words having semantic correlations, such as the third PC of blogs and the fourth PC of news (all contain “decrease/increase”), due to the biased word distribution of PPDB.

Although blogs and news corpora are expected to have different word distributions, they share the stylistic variation patterns mentioned above. One key difference between the principal components discovered in these two embedding spaces can be found in the second and third principal components of the news corpus, where “base (verb)  $\leftrightarrow$  present participle” is a dominant pattern, while it cannot be found in the top principal components of the blogs corpus.

Overall, this manual inspection suggests that the principal components do capture information that is relevant to style variations, even if they do not directly align to clear-cut style dimensions. Identifying how many top PCs are style-related (i.e., form a style subspace) is subjective and difficult. Therefore, we now turn to a quantitative evaluation.

### 3.3 Extrinsic Evaluation: Lexical Formality Scoring

We evaluate the usefulness of the latent dimensions discovered in Section 3.2 on a lexical formality prediction task. If the dimensions discovered are relevant to

style, they should help predict formality.

### 3.3.1 Identifying a Style Subspace

#### 3.3.1.1 Experimental Set-Up

**Task** Following [Brooke et al. \(2010\)](#), we use a list of 399 synonym pairs from a writing manual — *Choose the Right Word* (CTRW) ([Hayakawa, 1994](#)) — to evaluate the formality model. Given a pair of words, such as “hurry” vs. “expedite”, the task is to predict which is the more formal of the two.

**Ranking method** The predictions are made by linear *Support Vector Machine* (SVM, [Cortes and Vapnik, 1995](#)) classifiers (similar to the method proposed by [Brooke and Hirst \(2014\)](#)). They are trained on 105 formal seed words and 138 informal seed words used by [Brooke et al. \(2010\)](#). Each word is represented by a feature vector in word2vec spaces or their subspaces. When ranking two words, we actually compare their distances to the separating hyperplane, i.e.,  $\mathbf{w} \cdot \mathbf{e} - \mathbf{b}$ , where  $\mathbf{w}$ ,  $\mathbf{e}$  and  $\mathbf{b}$  are weight, embedding and bias.

**Embedding spaces** We first train word2vec (W2V) models on the blogs corpus with different vector space sizes (dimensionality=1–10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500). We then fix the vector space size of word2vec models to 300 since it provides a large enough original vector space and is a routinely used setting. All subspaces are extracted from these 300-dimensional original spaces.

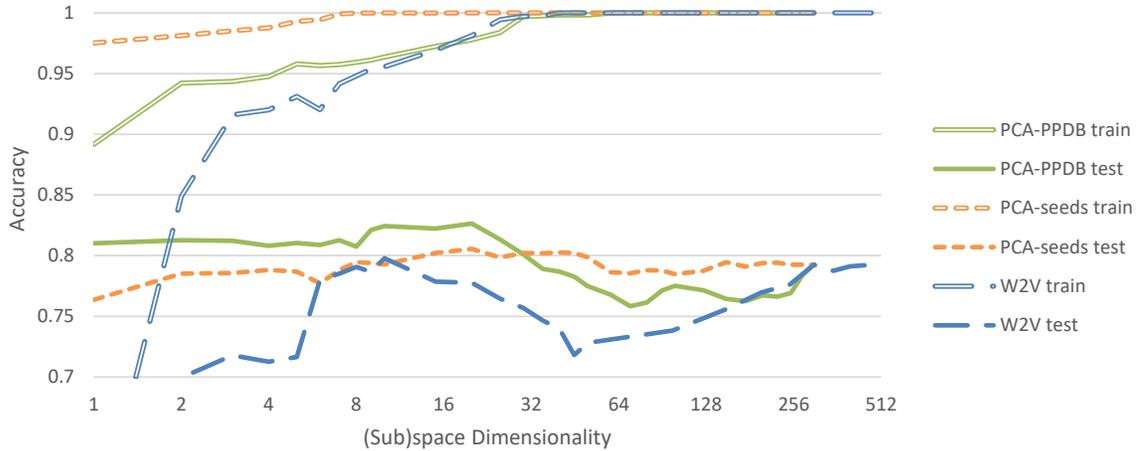


Figure 3.1: Train accuracy of formal/informal words classification and test accuracy of CTRW word-pair ranking vs. the (sub)space dimensionality. An SVM-based formality model achieved the best test performance on subspaces identified by PCA on PPDB data.

**Style subspaces** Next, we identify style subspaces (i.e., top PCs) using the PCA method introduced in Section 3.1. We examine every possible subspace size in the range of  $[1, 300]$  and denote this method as **PCA-PPDB**.

For comparison, we also train PCA subspaces using the seed words (**PCA-seeds**). Since seed words are not paraphrases, the PCA model is simply applied on word vectors. This method is based on the assumption that representative formal and informal words principally vary along the direction of formality.

### 3.3.1.2 Results

As illustrated in Figure 3.1, **\*\*\* train** indicates the training accuracy of SVM classifiers while **\*\*\* test** indicates the CTRW-pairs test accuracy.

The test accuracy of the W2V curve has two peaks when dimensionality=10 (accuracy=0.798) and dimensionality=300 (accuracy=0.792). Considering the near-

monotonicity of the training accuracy curve, we attribute the trough around dimensionality=45 to over-fitting (increasing number of features) while we attribute the rebound after that to more formality-related dimensions introduced.

Recall that we fix the original spaces to 300 dimensions. The accuracy curve provides another reason to choose this number: 300-dimensional original spaces can model formality well by itself and the performance converges when  $\text{dim} \geq 300$ .

Comparing `PCA-PPDB test` and `W2V test`, we can observe a clear advantage of using subspaces that capture latent lexical variations. Even a single first principal dimension surpassed original word2vec models of any size, including the full 300-dimensional space which yielded a test accuracy of 0.792. Further improvements were achieved when 9th-21st principal dimensions were introduced (max accuracy=0.826) — go back to Table 3.1, we can notice additional clues of formality variations from the 9th PC.

The accuracy curves of `PCA-seeds` indicate that this model can fit the training set better with fewer dimensions than the PPDB-based model but does not generalize as well to unseen test data. However, `PCA-seeds` still surpassed original word2vec models of any size.

### 3.3.2 SVM-Based Ranking vs. Other Formality Models

We have discussed the effectiveness of modeling formality using a subspace of small size (one for good performance and  $\sim 20$  for best performance). All analyses so far are based on a linear SVM, but can other sophisticated methods perform even

better on the style-embedded subspaces?

### 3.3.2.1 Formality Models

We compare SVM with state-of-the-art lexical formality models based on vector space models, such as SimDiff (Brooke et al., 2010) and Densifier (Rothe et al., 2016). Suppose each word  $w$  is represented as a  $k$ -dimensional vector  $\mathbf{e}_w$ .

SimDiff scores the formality of a word  $w$  by comparing its meaning to that of seed words of known formality using cosine similarity (Turney and Littman, 2003). Intuitively,  $w$  is more likely formal if it is semantically closer to formal seed words than to informal seed words.

Formally, given a formal word set  $S_F$  and an informal word set  $S_I$ , SimDiff scores a word  $w$  by calculating the difference between the similarity of  $w$  and each of these sets:

$$\text{score}(w) = \frac{1}{|S_F|} \sum_{v \in S_F} \cos(\mathbf{e}_w, \mathbf{e}_v) - \frac{1}{|S_I|} \sum_{v \in S_I} \cos(\mathbf{e}_w, \mathbf{e}_v). \quad (3.1)$$

While Brooke et al. (2010) use cosine to measure the similarity in *Latent Semantic Analysis* (LSA, Dumais et al., 1988; Deerwester et al., 1990) spaces, we replace it with dot product (i.e.,  $\mathbf{e}_w \cdot \mathbf{e}_v$ ) because it yields better results with word2vec embeddings on our test set.

Further manipulations such as score de-biasing and normalization are also introduced by Brooke et al. (2010), but they do not affect formality rankings examined by our evaluation.

**Densifier** is a supervised learning algorithm that transforms word embeddings into pre-defined dense orthogonal dimensions such as sentiment and concreteness. Under the formality ranking scenario, it optimizes a formality dimension  $\mathbf{d}$  (transition vector) that aims at separating words in  $S_F$  and words in  $S_I$  (i.e.,  $S_F \times S_I$ ), and grouping words in the same set (i.e.,  $S_F^2 \cup S_I^2$ ):

$$\arg \min_{\mathbf{d}} \sum_{(u,v) \in S_F^2 \cup S_I^2} \|\mathbf{d} \cdot (\mathbf{e}_u - \mathbf{e}_v)\| - \sum_{(u,v) \in S_F \times S_I} \mathbf{d} \cdot (\mathbf{e}_u - \mathbf{e}_v). \quad (3.2)$$

Note that the second term in this objective is equivalent to  $\arg \max_{\mathbf{d}} \sum_{(u,v) \in S_F \times S_I} \mathbf{d} \cdot \mathbf{e}_u + |\mathbf{d} \cdot \mathbf{e}_v|$ , which is similar to the objective of acquiring the first component  $\mathbf{d}$  from all data using PCA:  $\arg \max_{\mathbf{d}} \sum_{v \in S_F \cup S_I} (\mathbf{d} \cdot \mathbf{e}_v)^2$ .

The word formality can be simply assigned as the dot product of  $\mathbf{d}$  and  $\mathbf{e}_w$ :

$$\text{score}(w) = \mathbf{d} \cdot \mathbf{e}_w. \quad (3.3)$$

### 3.3.2.2 Results

All three formality scoring models (i.e., linear SVM, **SimDiff** and **Densifier**) are applied to subspaces extracted from 300-dimensional word2vec spaces using PCA on PPDB data. Figure 3.2 shows that these three models achieve nearly identical accuracy on subspaces with size smaller than 28.<sup>4</sup> Furthermore, we also compare the formality directions discovered by a linear SVM (coefficient  $\mathbf{w}$ ) and a **Densifier** (transition vector  $\mathbf{d}$ ). For any dimensionality, the cosine similarity between them is larger than 0.8. It is even larger than 0.9 when  $\text{dim} \geq 21$ . These suggest that the choice of ranking models has marginal impact, therefore identifying the style

---

<sup>4</sup>SVM could also have similar accuracy curve after dimension=28 if an RBF kernel was used.

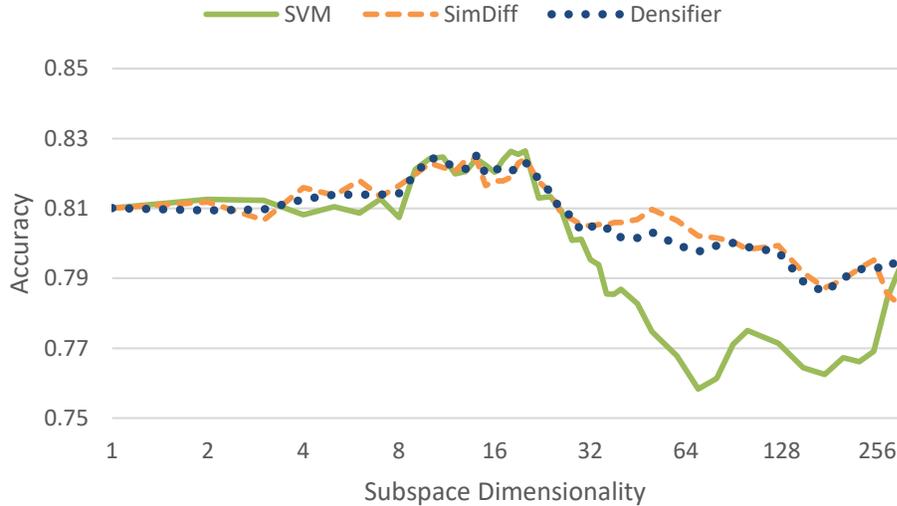


Figure 3.2: Test accuracy of CTRW word-pair ranking vs. the subspace dimensionality. All formality models achieved similar performance on subspaces of size 9-21 identified by PCA-PPDB.

subspace plays a more critical role in modeling formality.

### 3.3.3 Error Analysis

Identified subspaces capture formality decently in terms of ranking lexical formality — as high as 0.826 accuracy in the CTRW dataset (based on the best performing model, i.e., a linear SVM trained on a 20-dimensional subspace identified by PCA-PPDB). The question then arises: what types of errors contribute to the incorrect predictions?

Top (mis-)predicted CTRW word pairs are listed in Table 3.2, where  $s_i$  is the SVM (formality) score for word  $w_i$ .  $w_2$  is supposed to be more formal than  $w_1$ .

One category of errors roots in the mechanism of vector space models such as word2vec: they are all based on word co-occurrence patterns, which sometimes introduce unwanted biases. For example, “crony” itself is an informal synonym of

$w_1$	$w_2$	$s_1$	$s_2$	$s_2 - s_1$
Incorrect Examples				
crony	friend ††	0.667	-1.414	-2.081
conceit	vanity †	1.107	-0.697	-1.804
present †	gift	1.017	-0.732	-1.749
shiv	knife †	0.681	-0.863	-1.543
quotation	quote †	0.910	-0.594	-1.504
frighten	scare †	0.157	-1.244	-1.400
phony	fake †	0.237	-1.100	-1.337
parched	dehydrated †	0.173	-1.035	-1.209
punish †	chasen	0.260	-0.697	-0.956
penetrating †	perspicacious	1.527	0.644	-0.883
Correct Examples				
grill †	interrogate	-1.370	1.212	2.581
excuse †	remit	-0.608	2.001	2.609
gardening ††	tillage	-0.846	1.795	2.641
get ††	obtain	-1.435	1.296	2.731
hurry †	expedite	-1.632	1.174	2.806
catch ††	apprehend	-1.443	1.381	2.824
watch †	observe	-1.628	1.264	2.892
loud ††	clamorous	-1.304	1.819	3.123
quote ††	adduce	-0.594	2.529	3.123
beach ††	littoral	-1.116	2.143	3.259

Table 3.2: Top (mis-)predicted CTRW word pairs, where  $s_i$  is the SVM (formality) score for word  $w_i$ .  $w_2$  is supposed to be more formal than  $w_1$ . † This word is more frequent than the other in a pair according to the blogs corpus. (†/ ††/ ††† means at least 10/100/1000 times more.)

“friend” in our dataset. However, “crony capitalism” is a tightly glued economy term. For comparison, the formality score of “capitalism” is 0.966, which is very close to 0.667 of “crony”.

Ambiguity is another key factor that influences the formality scoring based on vector space models. [Arora et al. \(2018\)](#) pointed out that in the vector space, a word having multiple meanings lies in middle of its senses. Consequently, its formality score is also controlled by all its senses. We can find many ambiguous words in

the list of incorrect examples, such as “vanity” (clothing store, singer), “present”, “shiv” (Hindu god), “parched” (film), “chasen” (surname, band), etc.

Last but not least, word frequency is a strong signal of predicting formality, but it also sometimes misleads predictions. We use word frequencies in the blogs corpus to rank CTRW word pairs and got an accuracy as high as 0.771 (by arguably treating more frequent as less formal). Frequency information is not designed to be embedded into word2vec models, but it still can be partially reconstructed (Rothe et al., 2016). Projecting to the top (in)correct examples, a † symbol is placed behind the more frequent word in a pair. We can observe that top correctly ranked pairs follow the more-frequent-less-formal rule. However, this rule also biases the prediction to some incorrectly ranked pairs.

In a nutshell, formality models based on vector space models suffer from the limitation that a word representation is affected by word association, word sense and word frequency.

### 3.4 Summary

We presented an approach to discovering stylistic variations in distributional vector spaces using lexical paraphrases. Qualitative analysis suggested that the principal components discovered by PCA indeed captured variations related to style. Evaluation of a formality prediction task demonstrated the benefits of the induced subspace to detect style variations. We also compared the impact of different factors in identifying style-relevant dimensions such as the training data for PCA, the di-

mensionality of subspaces, and the nature of prediction methods. Finally, the error analysis indicated some intrinsic limitation of comparing style (formality) based on vector space models.

## Chapter 4: Reranking-Based Formality-Sensitive Machine Translation

Given formality annotations derived from modeling stylistic variations (Chapter 3), we are now able to control the formality of machine translation output. We introduce a new task for this purpose: *Formality-Sensitive Machine Translation* (FSMT). In addition to the input text in the source language, an FSMT system takes the desired formality for the output as input. This formality can be seen as approximating the intended audience of the translation. For example, the French sentence “Bonne idée, mais elle ne convient pas ici.” could be translated to “Good idea but it doesn’t fit here.”, which is informal because it elides the subject and uses contractions and chained clauses. It could also be translated more formally to “This is a helpful idea. However, it is not suitable for this purpose.”, which is grammatically complete and uses more formal and precise terms.

Our goal is to obtain a single MT system trained on diverse data which can adaptively produce output for a range of styles. By contrast, building multiple formality-specific systems is less flexible. To this end, we implement the initial FSMT system by  $n$ -best reranking — translation hypotheses matching desired formality level are promoted. This model is based on a standard PBMT architecture.

We first select a lexical formality model providing the most accurate scores on intrinsic sentential formality datasets. We then turn to machine translation and show that a lexical formality model can have a positive impact when used to control the formality of machine translation output. When the expected formality matches the reference, we obtain improvement of translation quality evaluated by an automatic metric (i.e., BLEU). A human assessment also verifies the effectiveness of our system in generating translations at diverse levels of formality.

## 4.1 Formality Modeling

The FSMT system requires quantifying the formality level of a sentence. Following prior work, we define sentence-level formality based on lexical formality scores (Brooke et al., 2010; Pavlick and Nenkova, 2015). We conduct an empirical comparison of existing techniques that can be adapted as lexical formality models, and introduce a sentence-level formality scheme based on the weighted average.<sup>1</sup>

### 4.1.1 Lexical Formality

#### 4.1.1.1 Models Based on Word Representations

We have discussed some prominent existing lexical formality models in Chapter 3 (Section 3.3), such as SVM (Brooke and Hirst, 2014), SimDiff (Brooke et al., 2010) and Densifier (Rothe et al., 2016).

Turning scores generated by different models into a unified scale requires fur-

---

<sup>1</sup>Code is available at <https://github.com/xingniu/computational-stylistic-variations>.

ther manipulation. A neutral word  $r$  has to be manually selected to anchor the midpoint of the formality score range. In other words, the final formality score for  $r$  is enforced to be zero:

$$\text{Formality}(w) = \frac{\text{score}(w) - \text{score}(r)}{\text{normalizer}(w, r)}. \quad (4.1)$$

The neutral word is typically selected from function words. We select “at” because it appears in nearly every document and appears with nearly equivalent probabilities in formal and informal corpora ( $S_F$  and  $S_I$ ). Finally, a normalizer which is maximized among the whole vocabulary ensures that scores cover the entire  $[-1, 1]$  range:

$$\text{normalizer}(w) = \begin{cases} \max_{v \in S_F} (\text{score}(v) - \text{score}(r)), & \text{if } \text{score}(w) - \text{score}(r) \geq 0 \\ \max_{v \in S_I} |\text{score}(v) - \text{score}(r)|, & \text{if } \text{score}(w) - \text{score}(r) < 0 \end{cases}. \quad (4.2)$$

In addition to **Densifier** which identifies a one-dimensional subspace that captures formality within the original vector space, we also directly train a PCA model on word representations of all seeds and chose the top principle component as the formality dimension.

#### 4.1.1.2 Models Based on Word Statistics

We also compare above models to a baseline that relies on unigram models to compare word statistics in corpora representative of formal vs. informal language (Pavlick and Nenkova, 2015). This model requires examples of formal and informal language and maps a word  $w$  to a continuous score:

$$\text{Formality}(w) = \log \frac{P(w | \text{FM})}{P(w | \text{FM} + \text{IFM})}, \quad (4.3)$$

where FM is the formal language corpus such as government documents, and IFM is the informal text such as telephone conversation transcripts. Word probabilities are estimated by unigram language models.

We modify this ratio to obtain scores that can be interpreted and used more easily. First, an adjusted ratio is defined as

$$r(w) = \text{sign}(c) \cdot \left[ \left( \frac{P(w | \text{FM})}{P(w | \text{IFM})} \right)^{\text{sign}(c)} - 1 \right], \quad (4.4)$$

where  $\text{sign}(c)$  extracts the sign of  $c = P(w | \text{FM}) - P(w | \text{IFM})$  and makes this function rotationally symmetric. The  $-1$  term aims at centering neutral words which have the same probabilities in both stylistic directions. The word count of  $w$  is smoothed to 0.1 if  $w$  is not in FM or IFM. Then, a simple sigmoid function with parameter  $\alpha$  can normalize the ratio to  $[-1, 1]$ :

$$\text{Formality}(w) = \frac{r(w)}{\alpha + |r(w)|}. \quad (4.5)$$

The normalization function is monotone so that the rankings obtained with the original formality score (in Equation 4.3) are retained, but it can distort the score density by tuning  $\alpha$ .<sup>2</sup> This model is denoted as **ProbRatio**.

#### 4.1.2 From Word to Sentence Formality

While previous work scored longer text by averaging word scores (Brooke and Hirst, 2014; Pavlick and Nenkova, 2015), we propose a weighted average scheme for word sequences  $W$  to downgrade the formality contribution of neutral words:

$$\text{Formality}(W) = \frac{\sum_{w_i \in W} |\text{Formality}(w_i)| \cdot \text{Formality}(w_i)}{\sum_{w_i \in W} |\text{Formality}(w_i)|}, \quad (4.6)$$

---

<sup>2</sup> $\alpha$  is set to 0.5 in our experiments.

where  $\text{Formality}(w)$  can be any of the lexical formality scores defined above.<sup>3</sup>

### 4.1.3 Evaluation

In order to choose an appropriate method for annotating MT data, we evaluate the formality models at the sentence level. [Lahiri \(2015\)](#) and [Pavlick and Tetreault \(2016\)](#) collect 5-way human scores for 11,263 sentences in the genres of blog, email, answers and news. Following [Pavlick and Tetreault \(2016\)](#), we average human scores for each sentence as the gold standard. We evaluate according to the root-mean-square error (RMSE) after re-scaling manual scores to  $[-1, 1]$ . RMSE takes into account the actual value of the formality score (cf. the correlation) and magnifies large errors (cf. the mean absolute error). It is arguably a more useful indicator of performance given our goal of using the formality score in downstream applications.

A large mixed-topic corpus is required to train vector space models. As in [Chapter 3](#), we use the ICWSM 2009 Spinn3r dataset (English tier-1) which consists of 1.6 billion words in 7.5 million English blogs ([Burton et al., 2009](#)). We also compare the term-document association model *Latent Semantic Analysis* (LSA, [Dumais et al., 1988](#); [Deerwester et al., 1990](#)) and the term-term association model word2vec (W2V, [Mikolov et al., 2013](#)). We use the same 105 formal seeds and 138 informal seeds as [Brooke et al. \(2010\)](#).

Following [Brooke et al. \(2010\)](#), to achieve best performance, we use a small dimensionality (i.e., 10) for training LSA and W2V. To achieve better performance, we normalize the LSA word vectors to make them have a unit length.

---

<sup>3</sup>The weighted average performs better than the standard average in preliminary experiments.

	LSA	W2V
SimDiff	0.353	0.404
SVM	0.361	0.424
PCA	0.352	0.390
Densifier	0.350	0.413
ProbRatio	0.395	

Table 4.1: Sentence-level formality quantifying evaluation (RMSE) among different models with different vector spaces.

**ProbRatio** requires language examples of diverse formality. Conversational transcripts are generally considered as casual text, so we concatenate corpora such as Fisher (Cieri et al., 2004), Switchboard (Godfrey et al., 1992), SBCSAE,<sup>4</sup> CallHome,<sup>5</sup> CallFriend,<sup>6</sup> BOLT SMS/Chat (Song et al., 2014) and NPS Chatroom (Forsyth and Martell, 2007). As the formal counterpart, we extract comparable size of English text from Europarl (Koehn, 2005). This results in 30 Million tokens of formal corpora (1.1M segments) and 29 Million tokens of informal corpora (2.7M segments).

Table 4.1 shows that LSA-based methods perform best on sentence-level evaluations. LSA captures term-document associations. At the sentence-level, such associations might help capture topic words that are effective indicators of formality even if they do not represent stylistic variations. W2V co-occurrence is based on a narrow context window, and thus might not capture topic information as term-document co-occurrence can. So we select **Densifier-LSA** as a representative for our FSMT system.

<sup>4</sup><https://www.linguistics.ucsb.edu/research/santa-barbara-corpus>

<sup>5</sup><https://catalog.ldc.upenn.edu/LDC97S42>

<sup>6</sup><https://talkbank.org/access/CABank/CallFriend/>

## 4.2 Formality-Sensitive Machine Translation

FSMT takes two inputs: text in the source language to be translated (i.e.  $\mathbf{X}$ ) and a desired formality level capturing the intended audience of the translation (i.e.,  $\ell$ ). An FSMT model with parameters  $\theta$  aims at finding the most probable translation  $\hat{\mathbf{Y}}$ , i.e.

$$\hat{\mathbf{Y}} = \arg \max_{\mathbf{Y}_\ell} P(\mathbf{Y}_\ell | \mathbf{X}, \ell; \theta). \quad (4.7)$$

We propose to implement FSMT as  $n$ -best reranking within a standard PBMT architecture and therefore introduce a formality-scoring feature for reranking. For each English translation hypothesis  $\mathbf{Y}$ , given the formality level  $\ell$  as a parameter:

$$h(\mathbf{e}; \ell) = |\text{Formality}(\mathbf{Y}) - \ell| \quad (4.8)$$

where  $\text{Formality}(\mathbf{Y})$  is the sentence-level formality score for  $\mathbf{Y}$ .

This formality feature  $h(\mathbf{Y}; \ell)$ , along with standard model features, is fed into a standard reranking model. When training the reranking model, the parameter  $\ell$  is set to the actual formality score of the reference translation for each instance. At test time,  $\ell$  is provided by the user. The re-scoring weights help promote candidate sentences whose formality scores approach the expected level.

### 4.2.1 Experimental Set-Up

**Task and Data** We evaluate this approach on a French to English translation task. Two parallel French-English corpora are used: (1) MultiUN (Eisele and Chen, 2010), which is extracted from the United Nations website, and can be considered

to be formal text; (2) OpenSubtitles2016 (Lison and Tiedemann, 2016), which is extracted from movie and television subtitles, covers a wider spectrum of styles, but overall tends to be informal since it primarily contains transcripts of conversations. Each parallel corpus is split into a training set (100M English tokens), a tuning set (2.5K segments) and a test set (5K segments). Two corpora are then concatenated, such that training, tuning and test sets all contain a diversity of styles.

**MT Set-Up** The Moses (Koehn et al., 2007) toolkit is used to build our PBMT system. We follow the standard training pipeline with default parameters.<sup>7</sup> Word alignments are generated using `fast_align` (Dyer et al., 2013), and symmetrized using the *grow-diag-final-and* heuristic. We use 4-gram language models, trained using KenLM (Heafield, 2011). Model weights are tuned using batch MIRA (Cherry and Foster, 2012).

We use constant size  $n=1000$  for  $n$ -best lists in all experiments. The reranking is a log-linear model trained using batch MIRA.<sup>8</sup> We report results averaged over five random tuning re-starts to compensate for tuning noise (Clark et al., 2011).

**FSMT** In order to evaluate the impact of different input formality (e.g., low/neutral/high) on translation quality, ideally, we would like to have three human reference translations with different formality for each source sentence. Since such references are not available, we construct three sets of test data where instances are divided according to the formality level of the available reference translation. The sentential

---

<sup>7</sup><http://www.statmt.org/moses/?n=Moses.Baseline>

<sup>8</sup><https://github.com/moses-smt/mosesdecoder/tree/master/scripts/nbest-rescore>

Desired Formality	Informal test set	Neutral test set	Formal test set
None (baseline)	39.74	40.17	<b>47.97</b>
low	<b>40.27</b>	39.65	47.76
neutral	38.70	<b>40.46</b>	47.84
high	37.58	39.53	<b>47.97</b>

Table 4.2: Translation quality (BLEU scores) on informal/neutral/formal sentence sets given different desired formality levels ( $-0.4$ ,  $0.0$ ,  $0.4$ ). Best results with statistical significance are highlighted.

formality distribution in the tuning set shows that 97% of the reference translations fall into the range of  $[-0.6, 0.6]$ . We therefore set three formality bins — informal  $[-1, -0.2)$ , neutral formality  $[-0.2, 0.2]$ , and formal  $(0.2, 1]$  — and split the test set into these bins. We use **Densifier**-LSA and the training setting described above to translate the entire test set three times, with three different formality levels: low ( $-0.4$ ), neutral ( $0$ ) and high ( $0.4$ ).<sup>9</sup>

## 4.2.2 Automatic Evaluation

We first report standard automatic evaluation results using the BLEU score to compare FSMT output given different desired formality level on each bin (see Table 4.2).

The best BLEU scores for each formality level are obtained when the level of formality given as input to the MT system matches the nature of the text being translated, as can be seen in the scores along the diagonal in Table 4.2. Comparing with the baseline system, which produces the top translation from each  $n$ -best list, translation quality improves by  $+0.5$  BLEU on informal text,  $+0.3$  BLEU on neutral

<sup>9</sup> $\pm 0.4$  yields best BLEU on the tuning set.

text, and remains constant on formal text. The impact increases with the distance to formal language. This can be explained by the fact that more formal sentences tend to be longer, and the impact of alternate lexical choice for a small number of words per sentence is smaller in longer sentences. In addition, the formal sentences are mostly drawn from UN data which is sufficiently different from the other registers in the heterogeneous training corpus that the informal examples do not affect baseline performance on formal data.

### 4.2.3 Human Assessment

Automatic evaluation is limited to comparing output to a single reference: lower BLEU scores conflate translation errors and stylistic mismatch. Therefore, we conduct a human study of the formality vs. the quality.

We conduct a manual evaluation of the output of our FSMT system taking low/high formality levels (-0.4/0.4) as parameters. 42 non-identical translation pairs are randomly selected and are annotated by 15 volunteers. For each pair of segments, an average of seven volunteers are asked to select the segment that would be more appropriate in a formal setting (e.g., a job interview) than in a casual setting (e.g., chatting with friends). A default option of “N: neither of them is more formal or hard to say” is also available to annotators.

By majority voting, 20 pairs are annotated as “N”, indicating the two translations has no distinctions with respect to formality. For example, “A: how can they do this” vs. “B: how can they do that”. Given that the translations are restricted

to the  $n$ -best list, not all sentences could be translated into stylistically different language.

Of the remaining 21 pairs where annotators judge one output more formal than the other, in all but one case the translation produced by our FSMT system with high formality level parameter is judged to be more formal. Overall this indicates that our formality scoring and ranking procedure are effective.

To determine whether reranking based on formality might have a detrimental effect on quality, we also have annotators rate the fluency and adequacy of the segments. Inspired by [Graham et al. \(2013\)](#), annotators are first asked to assess fluency without a reference and separately adequacy with a reference. Both assessments use a sliding scale. Each segment is evaluated by an average of seven annotators. After rescaling the ratings into the  $[0, 1]$  range, we observe a 0.75 level of fluency for informal translations and 0.70 for formal ones. This slight difference fits our expectation that more casual language may feel more fluent while more formal language may feel more stilted. The adequacy ratings are 0.65 and 0.64 for informal and translations respectively, indicating that adjusting the level of formality had minimal effect on the adequacy of the result.

Some interesting examples are listed in [Table 4.3](#). Occasionally, the  $n$ -best list has no translation hypotheses with diverse formality, so the FSMT system drops necessary words, appends inessential words, or selects improper or even incorrect words to fit the target formality level. In the case of “how do you do”, the translation that is meant to be more casual is rated more formal. Because the system measures formality on the lexical level, it is not able to recognize this idiomatically formal

$\ell$	Examples	Comments
-0.4	... and then he <b>ran away</b> .	–
0.4	... and then he <b>escaped</b> .	<i>annotated as more formal</i>
-0.4	<b>anybody hurt</b> ?	–
0.4	is <b>someone wounded</b> ?	<i>annotated as more formal</i>
-0.4	he <b>shot himself</b> in the middle of it .	–
0.4	he <b>committed suicide</b> in the middle of it .	<i>annotated as more formal</i>
-0.4	to move <b>things</b> forward .	–
0.4	<b>in order</b> to move <b>the process</b> forward.	<i>annotated as more formal</i>
-0.4	i'm <b>a police officer</b> for <b>about</b> 40 years .	–
0.4	i'm <b>in the police force</b> of <b>approximately</b> 40 years .	<i>annotated as more formal</i>
-0.4	how do you do ?	<i>annotated as more formal</i>
0.4	how are you?	–
-0.4	oh , val , you should get the phone .	<i>missing words</i>
0.4	oh , val , you should have the phone (of pete) .	–
-0.4	<b>i believe</b> you've solved the case , lieutenant .	<i>additive words</i>
0.4	you solved the case , lieutenant .	–
REF	right by checkout .	
-0.4	right next to the <b>body</b> .	<i>incorrect word choice</i>
0.4	right next to the <b>fund</b> .	<i>incorrect word choice</i>

Table 4.3: Examples of variant translations to the same French source segment using low/high output formality levels (-0.4/0.4) as parameters. In general the variations lie on the direction of formality as expected, but occasionally translation errors occur.

phrase made up of words that are not inherently formal. Despite these issues, most of the output are formality-variant translations of the same French source segment, as expected.

### 4.3 Summary

We presented a PBMT-based framework for formality-sensitive machine translation, where a system produces translations at the desired formality level. Automatic and human evaluation showed the effectiveness of this system in controlling language formality without loss in translation quality. However, the space of possible outputs is limited to lexical changes and  $n$ -best translation hypotheses. We will

turn to using neural models to capture more context in the following chapters.

## Chapter 5: Bi-Directional Low-Resource Neural Machine Translation

Lexical formality models provide useful but imperfect estimation of sentential formality — lexical formality scores could be biased by word association, word sense and word frequency (as discussed in Section 3.3.3), and they are not able to characterize idiomatic phrases (as discussed in Section 4.2.3). By contrast, neural models provide a more promising approach to model formality of sentences.

Formality style transfer can be viewed as a low-resource MT problem given a limited number of parallel examples with diverse formality styles. NMT has become the new standard of MT as it consistently outperforms previous methods across domains and language pairs (Bojar et al., 2017; Cettolo et al., 2017). However, NMT systems still struggle compared to PBMT in low-resource or out-of-domain scenarios (Koehn and Knowles, 2017).

In this chapter, we research how to improve the translation quality of low-resource NMT independently of style by making better use of various sources of training data. In Section 5.1, we first propose a bi-directional NMT framework inspired by multi-task learning. It trains both directions of a language pair jointly with a single model. Joint training can leverage limited training data effectively via duplication. In Section 5.2, we further introduce a differentiable input reconstruction

loss to bi-directional NMT, aiming at exploiting the source side of parallel samples. This loss compares original inputs to reconstructed inputs, which are obtained by back-translating translation hypotheses into the input language.

## 5.1 Bi-Directional Models with Synthetic Parallel Data

A technique for overcoming a lack of data is multi-task learning, in which domain knowledge can be transferred between related tasks (Caruana, 1997). Johnson et al. (2017) apply the idea to multilingual NMT by concatenating parallel data of various language pairs and marking the source with the desired output language. The authors report promising results for translation between languages that have zero parallel data. This approach also dramatically reduces the complexity of deployment by packing multiple language pairs into a single model.

In many low-resource scenarios, parallel data is prohibitively expensive or otherwise impractical to collect, whereas monolingual data may be more abundant. NMT systems consist of one large neural network that performs full sequence-to-sequence translation. Trained end-to-end on parallel data, these models lack a direct avenue for incorporating monolingual data. Sennrich et al. (2016b) overcome this challenge by back-translating target monolingual data to produce *synthetic* parallel data that can be added to the training pool. While effective, back-translation introduces the significant cost of first building a reverse system.

We propose a novel combination of multilingual NMT and back-translation that trains both directions of a language pair jointly with a single model. Specifically,

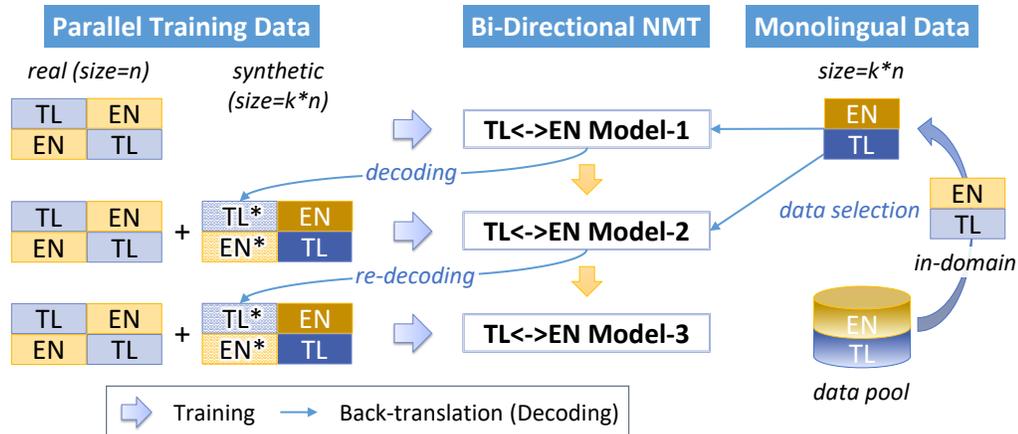


Figure 5.1: The framework of bi-directional NMT with synthetic parallel data. A bi-directional model (Model-1) is initialized on parallel data, and it translates select source and target monolingual data. Training is then continued on the augmented parallel data, leading to a cycle of improvement ( $\rightarrow$  Model-2  $\rightarrow$  Model-3).

we initialize a bi-directional model on parallel data and then use it to translate select source and target monolingual data. Training is then continued on the augmented parallel data, leading to a cycle of improvement. This approach (Figure 5.1) has several advantages:

- A single NMT model with standard architecture that performs all forward and backward translation during training.
- Training costs reduced significantly compared to uni-directional systems.
- Improvements in translating quality for low-resource languages, even over uni-directional systems with back-translation.
- Effectiveness in domain adaptation.

Via comprehensive experiments, we also contribute to best practices in selecting most suitable combinations of synthetic parallel data and choosing appropriate amount of monolingual data.

### 5.1.1 Approach

We introduce building bi-directional NMT with synthetic parallel data and present a strategy for selecting suitable monolingual data for back-translation.

#### 5.1.1.1 Bi-Directional NMT with Synthetic Parallel Data

We use the techniques described by [Johnson et al. \(2017\)](#) to build a multilingual model that combines forward and backward directions of a single language pair. To begin, we construct training data by swapping the source and target sentences of a parallel corpus and appending the swapped version to the original. We then add an artificial token to the beginning of each source sentence to mark the desired target language, such as `<2en>` for English. A standard NMT system can then be trained on the augmented dataset, which is naturally balanced between language directions.<sup>1</sup> A shared Byte-Pair Encoding (BPE) model is built on source and target data, alleviating the issue of unknown words and reducing the vocabulary to a smaller set of items shared across languages ([Sennrich et al., 2016c](#); [Johnson et al., 2017](#)). We further reduce model complexity by tying source and target word embeddings. The full training process significantly saves the total computing resources compared to training an individual model for each language direction.

Generating synthetic parallel data is straightforward with a bi-directional model: sentences from both source and target monolingual data can be translated to produce synthetic sentence pairs. Synthetic parallel data of the form `synthetic`

---

<sup>1</sup>[Johnson et al. \(2017\)](#) report the need to oversample when data is significantly unbalanced between language pairs.

→ `monolingual` can then be used in the forward direction, the backward direction, or both. Crucially, this approach leverages both source and target monolingual data while always placing the real data on the target side, eliminating the need for work-arounds such as freezing certain model parameters to avoid degradation from training on MT output (Zhang and Zong, 2016).

### 5.1.1.2 Monolingual Data Selection

Given the goal of improving a base bi-directional model, selecting ideal monolingual data for back-translation presents a significant challenge. Data too close to the original training data may not provide sufficient new information for the model. Conversely, data too far from the original data may be translated too poorly by the base model to be useful. We manage these risks by leveraging a standard pseudo in-domain data selection technique, cross-entropy difference (Moore and Lewis, 2010), to rank sentences from a general domain. Smaller cross-entropy difference indicates a sentence that is simultaneously more similar to the in-domain corpus (e.g., real parallel data) and less similar to the average of the general-domain monolingual corpus. This allows us to begin with “safe” monolingual data and incrementally expand to higher risk but potentially more informative data.

## 5.1.2 Experiments

In this section, we describe data, settings, and experimental methodology. We then present the results of comprehensive experiments designed to answer the

following questions: (1) How can synthetic data be most effectively used to improve translation quality? (2) Does the reduction in training time for bi-directional NMT come at the cost of lower translation quality? (3) Can we further improve training speed and translation quality training with incremental training and re-decoding? (4) How can we effectively choose monolingual training data? (5) How well does bi-directional NMT perform on domain adaptation?

### 5.1.2.1 Data

**Diverse Language Pairs:** We evaluate our approach on both high and low-resource data sets: German $\leftrightarrow$ English (DE $\leftrightarrow$ EN), Tagalog $\leftrightarrow$ English (TL $\leftrightarrow$ EN), and Swahili $\leftrightarrow$ English (SW $\leftrightarrow$ EN). Parallel and monolingual DE $\leftrightarrow$ EN data are provided by the WMT17 news translation task (Bojar et al., 2017). Parallel data for TL $\leftrightarrow$ EN and SW $\leftrightarrow$ EN contains a mixture of domains such as news and weblogs, and is provided as part of the IARPA MATERIAL program.<sup>2</sup> We split the shuffled original corpora into training, dev, and test sets, therefore they share a homogeneous n-gram distribution. For these low-resource pairs, TL and SW monolingual data are provided by the Common Crawl (Buck et al., 2014) while EN monolingual data is provided by the ICWSM 2009 Spinn3r blog dataset (tier-1, Burton et al., 2009).

**Diverse Domain Settings:** For WMT17 DE $\leftrightarrow$ EN, we choose news articles from 2016 (the closest year to the test set) as in-domain data for back-translation. For TL $\leftrightarrow$ EN and SW $\leftrightarrow$ EN, we identify in-domain and out-of-domain monolingual data and

---

<sup>2</sup><https://www.iarpa.gov/index.php/research-programs/material>

apply data selection to choose pseudo in-domain data (see Section 5.1.1.2). We use the training data as in-domain and either Common Crawl or ICWSM as out-of-domain. We also include a low-resource, long-distance domain adaptation task for these languages: training on News/Blog data and testing on Bible data. We split a parallel Bible corpus (Christodoulopoulos and Steedman, 2015) into sample, dev, and test sets, using the sample data as the in-domain seed for data selection.

**Preprocessing:** Following Hieber et al. (2017), we apply four pre-processing steps to parallel data: normalization, tokenization, sentence-filtering (length 80 cutoff), and joint source-target BPE with 50,000 operations (Sennrich et al., 2016c). Low-resource language pairs are also true-cased to reduce sparsity. BPE and true-casing models are rebuilt whenever the training data changes. Monolingual data for low-resource settings is filtered by retaining sentences longer than nine tokens. Itemized data statistics after preprocessing can be found in Table 5.1.

### 5.1.2.2 NMT Configuration

We use the attentional RNN encoder-decoder architecture implemented in the Sockeye toolkit (Hieber et al., 2017). Our translation model uses a bi-directional encoder with a single LSTM layer of size 512, multilayer perceptron attention with a layer size of 512, and word representations of size 512 (Bahdanau et al., 2015). We apply layer normalization (Ba et al., 2016) and tie source and target embedding parameters. We train using the Adam optimizer with a batch size of 64 sentences and checkpoint the model every 1000 updates (10,000 for DE↔EN) (Kingma and Ba,

Type	Dataset	# Sentences
High-resource: German↔English		
Training	Common Crawl + Europarl v7 + News Comm. v12	4,356,324
Dev	Newstest 2015+2016	5,168
Test	Newstest 2017	3,004
Mono-DE	News Crawl 2016	26,982,051
Mono-EN	News Crawl 2016	18,238,848
Low-resource: Tagalog↔English		
Training	News/Blog	50,705
Dev/Test	News/Blog	491/508
Dev/Test	Bible	500/500
Sample	Bible	61,195
Mono-TL	Common Crawl	26,788,048
Mono-EN	ICWSM 2009 blog	48,219,743
Low-resource: Swahili↔English		
Training	News/Blog	23,900
Dev/Test	News/Blog	491/509
Dev/Test	Bible	500/500
Sample	Bible	14,699
Mono-SW	Common Crawl	12,158,524
Mono-EN	ICWSM 2009 blog	48,219,743

Table 5.1: Data sizes of training, development, test, sample and monolingual sets. Sample data serves as the in-domain seed for data selection.

2015). Training stops after 8 checkpoints without improvement of perplexity on the development set. We decode with a beam size of 5.

For TL↔EN and SW↔EN, we add dropout to embeddings and RNNs of the encoder and decoder with probability 0.2. We also tie the output layer’s weight matrix with the source and target embeddings to reduce model size (Press and Wolf, 2017). The effectiveness of tying input and output target embeddings has been verified on several low-resource language pairs (Nguyen and Chiang, 2018).

For TL↔EN and SW↔EN, we train four randomly seeded models for each experi-

Uni-directional models							
ID	Training Data	TL→EN	EN→TL	SW→EN	EN→SW	DE→EN	EN→DE
U-1	L1→L2	31.99	31.28	32.60	39.98	29.51	23.01
U-2	L1→L2 + L1*→L2	<b>24.21</b>	<b>29.68</b>	<b>25.84</b>	<b>38.29</b>	<b>33.20</b>	<b>25.41</b>
U-3	L1→L2 + L1→L2*	22.13	27.14	24.89	36.53	30.89	23.72
U-4	L1→L2 + L1*→L2 + L1→L2*	23.38	29.31	25.33	37.46	33.01	25.05
Bi-directional models							
ID	L1=EN	L2=TL		L2=SW		L2=DE	
B-1	L1↔L2	32.72	31.66	33.59	39.12	28.84	22.45
B-2	L1↔L2 + L1*↔L2	32.90	32.33	33.70	39.68	29.17	24.45
B-3	L1↔L2 + L2*↔L1	32.71	31.10	33.70	39.17	31.71	21.71
B-4	L1↔L2 + L1*↔L2 + L2*↔L1	33.25	32.46	<b>34.23</b>	38.97	30.43	22.54
B-5	L1↔L2 + L1*→L2 + L2*→L1	<b>33.41</b>	<b>33.21</b>	34.11	<b>40.24</b>	<b>31.83</b>	<b>24.61</b>
B-5 <i>f</i>	L1↔L2 + L1*→L2 + L2*→L1	33.79	32.97	34.15	40.61	31.94	24.45
B-6 <i>f</i>	L1↔L2 + <u>L1*</u> →L2 + <u>L2*</u> →L1	<b>34.50</b>	<b>33.73</b>	<b>34.88</b>	<b>41.53</b>	<b>32.49</b>	<b>25.20</b>

Table 5.2: BLEU scores for uni-directional models (ID=U-*k*) and bi-directional NMT models (ID=B-*k*) trained on different combinations of real and synthetic parallel data. Models in B-5*f* are fine-tuned from base models in B-1. Best models in B-6*f* are fine-tuned from precedent models in B-5*f* and underscored synthetic data is re-decoded using precedent models. The highest score within each box is highlighted.

ment and combine them in a linear ensemble for decoding. For DE↔EN experiments, we train a single model and average the parameters of the best four checkpoints for decoding (Junczys-Dowmunt et al., 2016). We report case-insensitive BLEU with standard WMT tokenization.<sup>3</sup>

### 5.1.2.3 Uni-Directional NMT

We first evaluate the impact of synthetic parallel data on standard uni-directional NMT. Baseline systems trained on real parallel data are shown in row U-1 of Table 5.2.<sup>4</sup> In all tables, we use L1→L2 to indicate real parallel data where the source language is L1 and the target language is L2. Synthetic data is annotated by aster-

<sup>3</sup>We use the script <https://github.com/EdinburghNLP/nematus/blob/master/data/multi-bleu-detok.perl>

<sup>4</sup>Baseline BLEU scores are higher than expected on low-resource language pairs. We hypothesize that the data is homogeneous and easier to translate.

isks, such as  $L1^* \rightarrow L2$  indicating that  $L1^*$  is the synthetic back-translation of real monolingual data  $L2$ .

We always select monolingual data as an integer multiple of the amount of real parallel data  $n$ , i.e.,  $|L1 \rightarrow L2^*| = |L1^* \rightarrow L2| = kn$ . For  $DE \leftrightarrow EN$  models, we simply choose the top- $n$  sentences from shuffled News Crawl corpus. For all models of low-resource languages, we select the top- $3n$  sentences ranked by cross-entropy difference as described in Section 5.1.1.2. The choice of  $k$  is discussed in Section 5.1.2.6.

Shown in rows U-2 through U-4 of Table 5.2, we compare the results of incorporating different combinations of real and synthetic parallel data. Models trained on **only real data of target language** (i.e., in U-2) achieve better performance in BLEU than using other combinations. This is an expected result since translation quality is highly correlated with target language models. By contrast, standard back-translation is not effective for our low-resource scenarios. A significant drop ( $\sim 7$  BLEU points comparing U-1 and U-2 for  $TL/SW \rightarrow EN$ ) is observed when back-translating English. One possible reason is that the quality of the selected monolingual data, especially English, is not ideal. We will encounter this issue again when using bi-directional models with the same data in Section 5.1.2.4.

#### 5.1.2.4 Bi-Directional NMT

We map the same synthetic data combinations to bi-directional NMT, comparing against uni-directional models with respect to both translation quality and training time. Training bi-directional models requires doubling the training data

by adding a second copy of the parallel corpus where the source and target are swapped. We use the notation  $L1\leftrightarrow L2$  to represent the concatenation of  $L1\rightarrow L2$  and its swapped copy  $L2\rightarrow L1$  in Table 5.2.

Compared to independent models (i.e., U-1), the bi-directional  $DE\leftrightarrow EN$  model in B-1 is slightly worse (by  $\sim 0.6$  BLEU). These losses match observations by [Johnson et al. \(2017\)](#) on many-to-many multilingual NMT models. By contrast, most bi-directional low-resource models slightly outperform independent models. We hypothesize that in low-resource scenarios the neural model’s capacity is far from exhausted due to the redundancy in neural network parameters ([Denil et al., 2013](#)), and the benefit of training on twice as much data surpasses the detriment of confusing the model by mixing two languages.

We generate synthetic parallel data from the same monolingual data as in the uni-directional experiments. If we build training data symmetrically (i.e., B-2,3,4), back-translated sentences are distributed equally on the source and target sides, forcing the model to train on some amount of synthetic target data (i.e., MT output). For  $DE\leftrightarrow EN$  models, the best BLEU scores are achieved when synthetic training data is only present on the source side, while for low-resource models, the results are mixed. We see a particularly counter-intuitive result when using monolingual English data — no significant improvement (see B-3 for  $TL/SW\rightarrow EN$ ). As bi-directional models are able to leverage monolingual data of both languages, better results are achieved when combining all synthetic parallel data (see B-4 for  $TL/SW\rightarrow EN$ ). By further excluding potentially harmful target-side synthetic data (i.e., B-4  $\rightarrow$  B-5), the most unified and slim models achieve the best overall performance.

Model		TL→EN	EN→TL	SW→EN	EN→SW	DE→EN	EN→DE
Uni-directional	Baseline	76	78	63	66	41	48
	Synthetic	177	176	137	104	88	75
	TOTAL	507		371		252	
Bi-directional (fine-tuning)	Baseline		125		93		61
	Synthetic		285		218		113
	TOTAL	↓ 19%	410	↓ 14%	311	↓ 31%	174
	Synthetic	↓ 23%	219	↓ 44%	122	↓ 24%	86

Table 5.3: Number of checkpoints (=  $|\text{updates}|/1000$  for TL/SW↔EN or  $|\text{updates}|/10,000$  for DE↔EN) used by various NMT models. Bi-directional models (with fine-tuning) reduce training time significantly.

While the best bi-directional NMT models thus far (B-5) outperform the best uni-directional models (U-1) for low-resource language pairs, it is a struggle to match performance (U-2) in the high-resource DE↔EN scenario.

In terms of efficiency, bi-directional models consistently reduce the training time by 15-30% as shown in Table 5.3. Note that checkpoints are summed over all independent runs when ensemble decoding is used.<sup>5</sup>

### 5.1.2.5 Fine-Tuning and Re-Decoding

Training new NMT models from scratch after generating synthetic data is incredibly expensive, working against our goal of reducing the overall cost of deploying strong translation systems. Therefore, we continue training baseline models on augmented data as shown in B-5f of Table 5.2. These models achieve comparable translation quality to those trained from scratch (B-5) at a significantly reduced cost, i.e., 20-40% computing time reduction in the experiments illustrated in Table 5.3.

<sup>5</sup>The training time is proportional to the number of checkpoints.

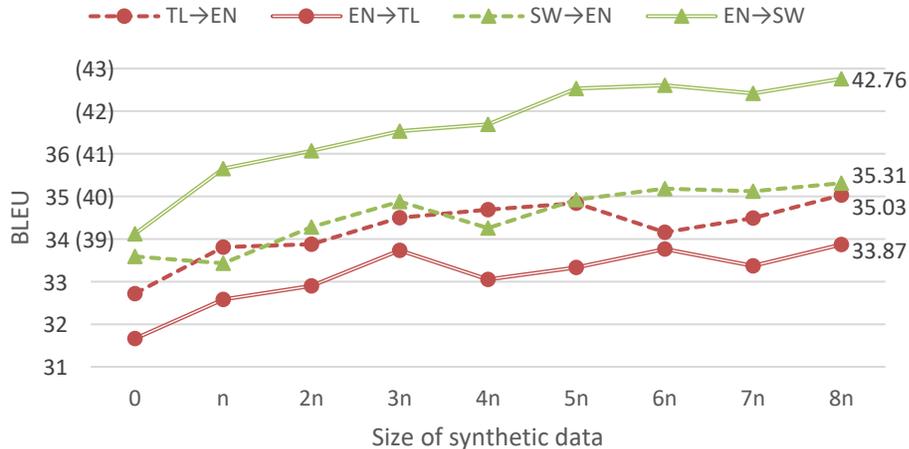


Figure 5.2: BLEU scores for four translation directions vs. the size of selected monolingual data.  $n$  in x-axis equals to the size of real parallel data.  $\text{EN} \rightarrow \text{SW}$  models use BLEU in parentheses in y-axis. Both language pairs tend to reach the plateau with more synthetic parallel data.

We also explore re-decoding the same monolingual data using improved models (Sennrich et al., 2016b). Underscored synthetic data in B-6f is re-decoded by models in B-5f, leading to the best results for all low-resource scenarios.

### 5.1.2.6 Size of Selected Monolingual Data

In our experiments, the optimal amount of monolingual data for constructing synthetic parallel data is task-dependent. Factors such as size and linguistic distribution of data and overlap between real parallel data, monolingual data, and test data can influence the effectiveness curve of synthetic data. We illustrate the impact of varying the size of selected monolingual data in our low-resource scenario. Shown in Figure 5.2, both language pairs tend to reach the plateau with more synthetic parallel data. The optimal point is a hyper-parameter that can be empirically determined on a tuning set.

ID	Training Data (L1=EN)	L2=TL		L2=SW	
		TL→EN	EN→TL	SW→EN	EN→SW
B-1	L1↔L2	11.03	10.17	6.56	3.80
B-5 <i>f</i>	L1↔L2 + L1*→L2 + L2*→L1	16.49	22.33	8.70	7.47
B-6 <i>f</i>	L1↔L2 + <u>L1</u> *→L2 + <u>L2</u> *→L1	<b>18.91</b>	<b>23.41</b>	<b>11.01</b>	<b>8.06</b>

Table 5.4: BLEU scores for bi-directional NMT models on Bible data. Models in B-5*f* are fine-tuned from baseline models in B-1. Highlighted best models in B-6*f* are fine-tuned from precedent models in B-5*f* and underscored synthetic data is re-decoded using precedent models. Baseline models are significantly improved in terms of BLEU.

### 5.1.2.7 Domain Adaptation

We evaluate the performance of using the same bi-directional NMT framework on a long-distance domain adaptation task: News/Blog to Bible. This task is particularly challenging because out-of-vocabulary (word type) rates of Bible test sets are as high as 30-45% when training on News/Blog. Significant linguistic differences also exist between modern and Biblical language use. The impact of this domain mismatch is demonstrated by the incredibly low BLEU scores of baseline News/Blog systems (Table 5.4, B-1). After fine-tuning baseline models on augmented parallel data (B-5*f*) and re-decoding (B-6*f*),<sup>6</sup> we see BLEU scores increase by 70-130%. Despite being based on extremely weak baseline performance, they still show the promise of our approach for domain adaptation.

<sup>6</sup>The concatenation of development sets from both News/Blog and Bible serves for validation.

## 5.2 Bi-Directional Differentiable Input Reconstruction

In Section 5.1, we improve low-resource NMT by duplicating parallel data and leveraging monolingual data. We hypothesize that the traditional training can be complemented by better leveraging limited training data. To this end, we propose a new training objective for this model by augmenting the standard translation cross-entropy loss with a **differentiable input reconstruction loss** to further exploit the source side of parallel samples.<sup>7</sup>

Input reconstruction is motivated by the idea of round-trip translation. Suppose sentence  $\mathbf{X}$  is translated forward to  $\hat{\mathbf{Y}}$  using model  $\theta_{XY}$  and then translated back to  $\hat{\mathbf{X}}$  using model  $\theta_{YX}$ , then  $\hat{\mathbf{Y}}$  is more likely to be a good translation if the distance between  $\hat{\mathbf{X}}$  and  $\mathbf{X}$  is small (Brislin, 1970). Prior work applied round-trip translation to monolingual examples and sampled the intermediate translation  $\hat{\mathbf{Y}}$  from a  $n$ -best list generated by model  $\theta_{XY}$  using beam search (Cheng et al., 2016; He et al., 2016). However, beam search is not differentiable which prevents back-propagating reconstruction errors to  $\theta_{XY}$ . As a result, reinforcement learning algorithms, or independent updates to  $\theta_{XY}$  and  $\theta_{YX}$  were required.

In this section, we focus on the problem of making input reconstruction differentiable to simplify training. In past work, Tu et al. (2017) addressed this issue by reconstructing source sentences from the decoder’s hidden states. However, this reconstruction task can be artificially easy if hidden states over-memorize the input. This approach also requires a separate auxiliary reconstructor, which introduces

---

<sup>7</sup>Implementation is available at <https://github.com/xingniu/sockeye/tree/naacl2019>.

additional parameters.

We propose instead to combine benefits from differentiable sampling and bi-directional NMT to obtain a compact model that can be trained end-to-end with back-propagation. Specifically,

- Translations are sampled using the Straight-Through Gumbel Softmax (STGS) estimator (Jang et al., 2017; Bengio et al., 2013), which allows back-propagating reconstruction errors.
- Our approach builds on the bi-directional NMT model, which improves low-resource translation by jointly modeling translation in both directions (e.g., Swahili  $\leftrightarrow$  English). A single bi-directional model is used as a translator and a reconstructor (i.e.,  $\theta_{XY} = \theta_{YX}$ ) without introducing more parameters.

Experiments show that our approach outperforms reconstruction from hidden states. It achieves consistent improvements across various low-resource language pairs and directions, showing its effectiveness in making better use of limited parallel data.

### 5.2.1 Approach

Recall that in our bi-directional model, the source sentence can be either  $\mathbf{X}$  or  $\mathbf{Y}$  and is respectively translated to  $\mathbf{Y}$  or  $\mathbf{X}$ . The language is marked by a tag (e.g., `<2en>`) at the beginning of each source sentence. To facilitate symmetric reconstruction, we also add language tags to target sentences. The training data corpus is then built by swapping the source and target sentences of a parallel corpus

and appending the swapped version to the original.

### 5.2.1.1 Bi-Directional Reconstruction

Our bi-directional model performs both forward translation and backward reconstruction. By contrast, uni-directional models require an auxiliary reconstruction module, which introduces additional parameters. This module can be either a decoder-based reconstructor (Tu et al., 2017; Wang et al., 2018a,b) or a reversed dual NMT model (Cheng et al., 2016; He et al., 2016; Wang et al., 2018c; Zhang et al., 2018).

Here the reconstructor, which shares the same parameter with the translator  $\text{MT}(\cdot)$ , can also be trained end-to-end by maximizing the log-likelihood of reconstructing  $\mathbf{X}$ :

$$\mathcal{L}_{RC} = \sum_{\mathbf{X}} \log P(\mathbf{X} | \text{MT}(\mathbf{X}; \boldsymbol{\theta}); \boldsymbol{\theta}). \quad (5.1)$$

Combining with the forward translation likelihood  $\mathcal{L}_{MT}$  in Equation 2.9, we use  $\mathcal{L}_{MT} + \mathcal{L}_{RC}$  as the final training objective for  $\mathbf{X} \rightarrow \mathbf{Y}$ . The dual  $\mathbf{Y} \rightarrow \mathbf{X}$  model is trained simultaneously by swapping the language direction in bi-directional NMT.

Reconstruction is reliable only with a model that produces reasonable base translations. Following prior work (Tu et al., 2017; He et al., 2016; Cheng et al., 2016), we pre-train a base model with  $\mathcal{L}_{MT}$  and fine-tune it with  $\mathcal{L}_{MT} + \mathcal{L}_{RC}$ .

### 5.2.1.2 Differentiable Sampling

We use differentiable sampling to side-step beam search and back-propagate error signals. We use the Gumbel-Max reparameterization trick (Maddison et al., 2014) to sample a translation token at each time step from the softmax distribution in Equation 2.8:

$$\mathbf{y}_t = \text{one-hot} \left( \arg \max_k (a(\mathbf{h}_t)_k + G_k) \right) \quad (5.2)$$

where  $a(\mathbf{h}_t) = \mathbf{W}\mathbf{h}_t + \mathbf{b}$  and  $G_k$  is i.i.d. and drawn from  $\text{Gumbel}(0, 1)$ .<sup>8</sup> We use scaled Gumbel with parameter  $\beta$ , i.e.,  $\text{Gumbel}(0, \beta)$ , to control the randomness. The sampling becomes deterministic (which is equivalent to greedy search) as  $\beta$  approaches 0.

Since  $\arg \max$  is not a differentiable operation, we approximate its gradient with the Straight-Through Gumbel Softmax (STGS) (Jang et al., 2017; Bengio et al., 2013):  $\nabla_{\theta} \mathbf{y}_t \approx \nabla_{\theta} \tilde{\mathbf{y}}_t$ , where

$$\tilde{\mathbf{y}}_t = \text{softmax} \left( (a(\mathbf{h}_t) + G) / \tau \right) \quad (5.3)$$

As  $\tau$  approaches 0, softmax is closer to arg max but training might be more unstable. While the STGS estimator is biased when  $\tau$  is large, it performs well in practice (Gu et al., 2018; Choi et al., 2018) and is sometimes faster and more effective than reinforcement learning (Havrylov and Titov, 2017).

To generate coherent intermediate translations, the decoder used for sampling only consumes its previously predicted  $\hat{\mathbf{Y}}_{<t}$ . This contrasts with the usual *teacher*

---

<sup>8</sup>i.e.,  $G_k = -\log(-\log(u_k))$  and  $u_k \sim \text{Uniform}(0, 1)$ .

*forcing* strategy (Williams and Zipser, 1989), which always feeds in the ground-truth previous tokens  $\mathbf{Y}_{<t}$  when predicting the current token  $\hat{\mathbf{y}}_t$ . With teacher forcing, the sequence concatenation  $[\mathbf{Y}_{<t}; \hat{\mathbf{y}}_t]$  is probably coherent at each time step, but the actual predicted sequence  $[\hat{\mathbf{Y}}_{<t}; \hat{\mathbf{y}}_t]$  would break the continuity.<sup>9</sup>

## 5.2.2 Experiments

### 5.2.2.1 Tasks and Data

We evaluate our approach on four low-resource language pairs. Parallel data for Swahili $\leftrightarrow$ English (SW $\leftrightarrow$ EN), Tagalog $\leftrightarrow$ English (TL $\leftrightarrow$ EN) and Somali $\leftrightarrow$ English (SO $\leftrightarrow$ EN) contains a mixture of domains such as news and weblogs and is collected from the IARPA MATERIAL program, the Global Voices parallel corpus<sup>10</sup>, Common Crawl (Smith et al., 2013), and the LORELEI Somali representative language pack (LDC2018T11). The test samples are extracted from the held-out ANALYSIS set of MATERIAL. Parallel Turkish $\leftrightarrow$ English (TR $\leftrightarrow$ EN) data is provided by the WMT news translation task (Bojar et al., 2018). We use pre-processed “corpus”, “newsdev2016”, “newstest2017” as training, development and test sets.<sup>11</sup>

As in Section 5.1, we apply normalization, tokenization, true-casing, joint source-target BPE with 32,000 operations (Senrich et al., 2016c) and sentence-filtering (length 80 cutoff) to parallel data.<sup>12</sup> Itemized data statistics after prepro-

---

<sup>9</sup>Sampling with teacher forcing yielded consistently worse BLEU than baselines in preliminary experiments.

<sup>10</sup><http://casmacat.eu/corpus/global-voices.html>

<sup>11</sup><http://data.statmt.org/wmt18/translation-task/preprocessed/>

<sup>12</sup>Less BPE operations are used in this section because a smaller vocabulary yields better low-resource performance.

# sent.	Training	Dev.	Test
SW↔EN	60,570	500	3,000
TL↔EN	70,703	704	3,000
SO↔EN	68,550	844	3,000
TR↔EN	207,021	1,001	3,007

Table 5.5: Experiments are conducted on four low-resource language pairs, in both translation directions.

cessing can be found in Table 5.5. We report case-insensitive BLEU with the WMT standard ‘13a’ tokenization using SacreBLEU (Post, 2018).

### 5.2.2.2 Model Configuration and Baseline

We build NMT models upon the attentional RNN encoder-decoder architecture (Bahdanau et al., 2015) implemented in the `Sockeye` toolkit (Hieber et al., 2017) with the same settings introduced in Section 5.1. Our translation model uses a bi-directional encoder with a single LSTM layer of size 512, multilayer perceptron attention with a layer size of 512, and word representations of size 512. We apply layer normalization (Ba et al., 2016) and add dropout to embeddings and RNNs (Gal and Ghahramani, 2016) with probability 0.2. We train using the Adam optimizer (Kingma and Ba, 2015) with a batch size of 48 sentences and we checkpoint the model every 1000 updates.<sup>13</sup> The learning rate for baseline models is initialized to 0.001 and reduced by 30% after 4 checkpoints without improvement of perplexity on the development set. Training stops after 10 checkpoints without improvement.

The bi-directional NMT model ties source and target embeddings to yield a

---

<sup>13</sup>Smaller batch size is used in this section to fit the GPU memory since the new loss enlarges the computational graph.

bilingual vector space. It also ties the output layer’s weights and embeddings to achieve better performance in low-resource scenarios (Press and Wolf, 2017; Nguyen and Chiang, 2018).

We train five randomly seeded bi-directional baseline models by optimizing the forward translation objective  $\mathcal{L}_{MT}$  and report the mean and standard deviation of test BLEU. We fine-tune baseline models with objective  $\mathcal{L}_{MT} + \mathcal{L}_{RC}$ , inheriting all settings except the learning rate which is re-initialized to 0.0001. Each randomly seeded model is fine-tuned independently, so we are able to report the standard deviation of  $\Delta$ BLEU.

### 5.2.2.3 Contrastive Reconstruction Model

We compare our approach with reconstruction from hidden states (HIDDEN). Following the best practice of Wang et al. (2018a), two reconstructors are used to take hidden states from both the encoder and the decoder. The corresponding two reconstruction losses and the canonical translation loss were originally uniformly weighted (i.e., 1, 1, 1), but we found that balancing the reconstruction and translation losses yields better results (i.e., 0.5, 0.5, 1) in preliminary experiments.<sup>14</sup>

We use the reconstructor exclusively to compute the reconstruction training loss. It has also been used to re-rank translation hypotheses in prior work, but Tu et al. (2017) showed in ablation studies that the gains from re-ranking are small compared to those from training.

---

<sup>14</sup>We observed around 0.2 BLEU gains for TR $\leftrightarrow$ EN tasks.

Model	EN→SW	SW→EN	EN→TL	TL→EN
Baseline	33.60 ± 0.14	30.70 ± 0.19	27.23 ± 0.11	32.15 ± 0.21
HIDDEN	33.41 ± 0.15	30.91 ± 0.19	27.43 ± 0.14	32.20 ± 0.35
Δ	<b>-0.19 ± 0.24</b>	<b>0.21 ± 0.14</b>	<b>0.19 ± 0.13</b>	0.04 ± 0.17
$\beta = 0$	33.92 ± 0.10	31.37 ± 0.18	27.65 ± 0.09	32.75 ± 0.32
Δ	<b>0.32 ± 0.12</b>	<b>0.66 ± 0.11</b>	<b>0.42 ± 0.16</b>	<b>0.59 ± 0.13</b>
$\beta = 0.5$	33.97 ± 0.08	31.39 ± 0.09	27.65 ± 0.10	32.65 ± 0.24
Δ	<b>0.37 ± 0.09</b>	<b>0.69 ± 0.11</b>	<b>0.42 ± 0.11</b>	<b>0.50 ± 0.08</b>

Model	EN→SO	SO→EN	EN→TR	TR→EN
Baseline	12.25 ± 0.08	20.80 ± 0.12	12.90 ± 0.04	15.32 ± 0.11
HIDDEN	12.30 ± 0.11	20.72 ± 0.16	12.77 ± 0.11	15.34 ± 0.10
Δ	0.05 ± 0.11	<b>-0.08 ± 0.12</b>	<b>-0.13 ± 0.13</b>	0.01 ± 0.07
$\beta = 0$	12.47 ± 0.08	21.14 ± 0.19	13.26 ± 0.07	15.60 ± 0.19
Δ	<b>0.22 ± 0.04</b>	<b>0.35 ± 0.15</b>	<b>0.36 ± 0.09</b>	<b>0.28 ± 0.11</b>
$\beta = 0.5$	12.48 ± 0.09	21.20 ± 0.14	13.16 ± 0.08	15.52 ± 0.07
Δ	<b>0.23 ± 0.03</b>	<b>0.41 ± 0.13</b>	<b>0.25 ± 0.09</b>	<b>0.19 ± 0.05</b>

Table 5.6: BLEU scores on eight translation directions. The numbers before and after ‘±’ are the mean and standard deviation over five randomly seeded models. Our proposed methods ( $\beta = 0/0.5$ ) achieve small but consistent improvements.  $\Delta$ BLEU scores are in bold if mean–std is above zero while in red if the mean is below zero.

#### 5.2.2.4 Results

Table 5.6 shows that our reconstruction approach achieves small but consistent BLEU improvements over the baseline on all eight tasks.<sup>15</sup>

We evaluate the impact of the Gumbel Softmax hyperparameters on the development set. We select  $\tau = 2$  and  $\beta = 0/0.5$  based on training stability and BLEU. Greedy search (i.e.,  $\beta = 0$ ) performs similarly as sampling with increased Gumbel noise (i.e., more random translation selection when  $\beta = 0.5$ ): increased randomness

<sup>15</sup>The improvements are significant with  $p < 0.01$ .

in sampling does not have a strong impact on BLEU, even though random sampling may approximate the data distribution better (Ott et al., 2018). We hypothesize that more random translation selection introduces lower quality samples and therefore noisier training signals. This is consistent with the observation that random sampling is less effective for back-translation in low-resource settings (Edunov et al., 2018).

Sampling-based reconstruction is effective even if there is moderate domain mismatch between the training and the test data, such as in the case that the word type out-of-vocabulary (OOV) rate of TR→EN is larger than 20%. Larger improvements can be achieved when the test data is closer to training examples. For example, the OOV rate of SW→EN is much smaller than the OOV rate of TR→EN and the former obtains higher  $\Delta$ BLEU.

Our approach yields more consistent results than reconstructing from hidden states. The latter fails to improve BLEU in more difficult cases, such as TR↔EN with high OOV rates. We observe extremely low training perplexity for HIDDEN compared with our proposed approach (Figure 5.3a). This suggests that HIDDEN yields representations that memorize the input rather than improve output representations.

Another advantage of our approach is that all parameters were jointly pre-trained, which results in more stable training behavior. By contrast, reconstructing from hidden states requires to initialize the reconstructors independently and suffers from unstable early training behavior (Figure 5.3).

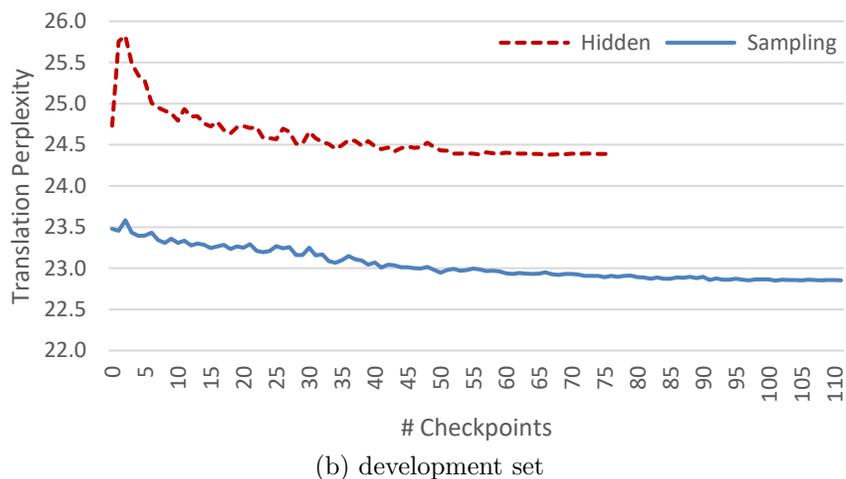
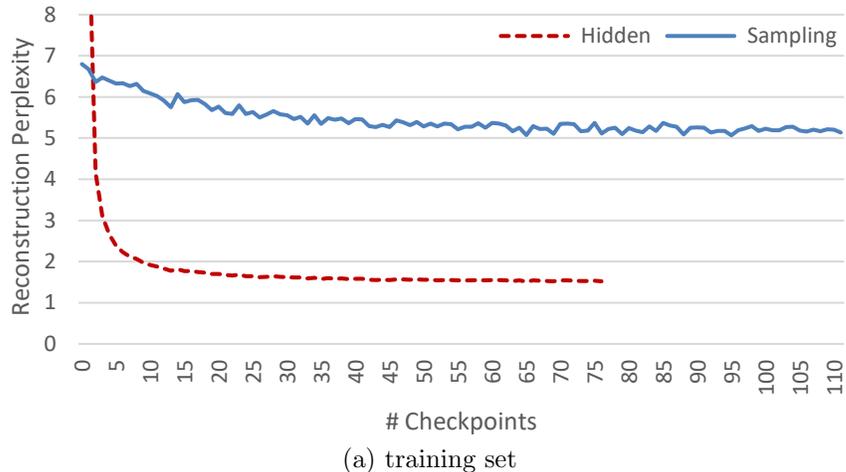


Figure 5.3: Training curves of perplexity on the training and the development sets for TR $\leftrightarrow$ EN. Reconstructing from hidden states (HIDDEN) and reconstructing from sampled translations ( $\beta = 0$ ) are compared. HIDDEN achieves extremely low training perplexity and suffers from unstable training during the early stage.

### 5.3 Summary

We introduced novel approaches to improve the translation quality of low-resource NMT by making better use of various sources of training data. We first presented the bi-directional NMT. This single model with a standard NMT architecture performs both forward and backward translation, allowing it to back-translate and incorporate any source or target monolingual data. By continuing training

on augmented parallel data, bi-directional NMT models consistently achieved improved translation quality, particularly in low-resource scenarios and cross-domain tasks. These models also reduced training and deployment costs significantly compared to standard uni-directional models used in iterative back-translation (Zhang et al., 2018; Hoang et al., 2018; Cotterell and Kreutzer, 2018). On top of the bi-directional NMT, we then studied reconstructing the input of NMT from its intermediate translations to better exploit training samples in low-resource settings. We used the Straight-Through Gumbel Softmax to build a fully differentiable reconstruction model that does not require any additional parameters. We empirically demonstrated that our approach is effective in low-resource scenarios.

## Chapter 6: Multi-Task Neural Formality Transfer and FSMT

Formality Transfer (FT) and Formality-Sensitive Machine Translation (FSMT) can both be framed as machine translation, but appropriate training examples are much harder to obtain than for traditional machine translation tasks. We hypothesize that FT and FSMT can benefit from being addressed jointly, by sharing information learned from two different types of supervision: sentence pairs in the same language that capture style difference, and translation pairs drawn from corpora of various styles.

In this Chapter, we first apply the bi-directional model from our low-resource NMT research (e.g., Chapter 5) to English FT tasks. It yields an elegant and unified model that transfers between formal and informal language. We then adopt the idea of multi-task learning by jointly training bi-directional formality transfer and machine translation. Training our model shares information from two distinct types of supervision: sentence pairs in the same language that capture formality difference, and translation pairs drawn from corpora of diverse formality. Designing this model requires addressing several questions: How can we effectively combine monolingual examples of formality transfer and bilingual examples of translation? What kind of bilingual examples are most useful for the joint task? Can our joint model learn

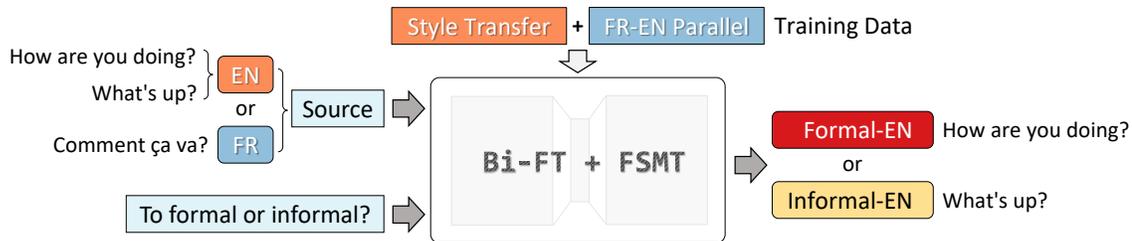


Figure 6.1: System overview: Our multi-task learning model can perform both bi-directional English formality transfer and translate French to English with desired formality. It is trained jointly on monolingual formality transfer data and bilingual translation data.

to perform FSMT without being explicitly trained on style-annotated translation examples? We explore these questions by conducting an empirical study on English FT and French-English FSMT, using both automatic and human evaluation.

The joint training yields a single model that performs both FT and FSMT (see Figure 6.1). The same model improves the state-of-the-art on the FT task and achieves competitive performance on FSMT without being explicitly trained on style-annotated translation examples.<sup>1</sup>

## 6.1 Approach

We describe our unified model for performing FT in both directions (Section 6.1.1), our FSMT model with side constraints (Section 6.1.2) and finally our multi-task learning model that jointly learns to perform FT and FSMT (Section 6.1.3). All models rely on the same NMT architecture: attentional recurrent sequence-to-sequence models.

<sup>1</sup>Data and scripts are available at <https://github.com/xingniu/multitask-ft-fsmt>.

### 6.1.1 Bi-Directional Formality Transfer

[Rao and Tetreault \(2018\)](#) used independent neural machine translation models for each formality transfer direction (`informal`→`formal` and `formal`→`informal`). Inspired by the bi-directional NMT for low-resource languages, we propose a unified model that can handle either direction — we concatenate the parallel data from the two directions of formality transfer and attach a tag to the beginning of each source sentence denoting the desired target formality level i.e., `<F>` for transferring to formal and `<I>` for transferring to informal. This enables our FT model to learn to transfer to the correct style via attending to the tag in the source embedding. We train an NMT model on this combined dataset. Since both the source and target sentences come from the same language, we encourage their representations to lie in the same distributional vector space by (1) building a shared Byte-Pair Encoding (BPE) model on source and target data ([Sennrich et al., 2016c](#)) and (2) tying source and target word embeddings ([Press and Wolf, 2017](#)).

### 6.1.2 Formality-Sensitive Machine Translation with Side Constraints

Inspired by [Sennrich et al. \(2016a\)](#), we use side constraints on parallel translation examples to control the output formality. At training time, this requires a tag that captures the formality of the target sentence for every sentence pair. Given the vast range of text variations that influence style, we cannot obtain tags using rules as for T-V pronoun distinctions ([Sennrich et al., 2016a](#)). Instead, we categorize French-English parallel data into formal vs. informal categories by comparing them

to the informal and formal English from the GYAFC corpus (Rao and Tetreault, 2018).

We adopt a data selection technique, *Cross-Entropy Difference* (CED, Moore and Lewis, 2010), to rank English sentences in the bilingual corpus by their relative distance to each style. First, we consider formal English as the target style and define  $CED(s) = H_{formal}(s) - H_{informal}(s)$ , where  $H_{formal}(s)$  is the cross-entropy between a sentence  $s$  and the formal language model. Smaller CED indicates an English sentence that is more similar to the formal English corpus and less similar to the informal English corpus. We rank English sentences by their CED scores and select the top  $N$  sentences (the choice of  $N$  is discussed in Section 6.4). Pairing these  $N$  English sentences with their parallel French source, we get the formal sample of our bilingual data. Similarly, we construct the informal sample using informal English as the target style. Finally, we combine the formal and the informal samples, attach the <F> and <I> tags to corresponding source French sentences (i.e., the bottom two rows of data in Figure 6.2a) and train an NMT model for our FSMT task.

### 6.1.3 Multi-Task Learning

We propose a multi-task learning model to jointly perform FT and FSMT using a many-to-one (i.e., multi-language to English) sequence to sequence model (Luong et al., 2016). Following Johnson et al. (2017), we implement this approach using shared encoders and decoders. This approach can use existing NMT architectures without modifications. To best incorporate side constraints at training time and the

<F>	Informal-EN	Formal-EN
<I>	Formal-EN	Informal-EN
<F>	FR	Formal-EN
<I>	FR	Informal-EN

<F>	Informal-EN	Formal-EN
<I>	Formal-EN	Informal-EN
	FR	Formal-EN
	FR	Informal-EN

(a) **MultiTask-tag-style**: formality tags on bilingual data + 2-style selection

(b) **MultiTask-style**: no formality tags on bilingual data + 2-style selection

<F>	Informal-EN	Formal-EN
<I>	Formal-EN	Informal-EN
	FR	EN

(c) **MultiTask-random**: no formality tags on bilingual data + random selection

Figure 6.2: The training data used for multi-task learning models. The bi-directional formality transfer data and the bilingual data (e.g., FR-EN) of equivalent size are always concatenated.

benefits of sharing representations for style and language, we explore three model designs.

**MultiTask-tag-style** is a straightforward combination of the transfer and translation models above. We hypothesize that using the bilingual parallel data where English is the target could enhance English FT in terms of target language modeling, especially when the bilingual data has similar topics and styles. We therefore combine equal sizes of formality tagged training data (selected as described in Section 6.1.2) from our FT and FSMT tasks in this configuration (Figure 6.2a).

**MultiTask-style** is designed to test whether formality tags for bilingual examples are necessary. We hypothesize that the knowledge of controlling the target formality for the FSMT task can be learned from the FT data since the source embeddings of formality tags are shared between the FT and the FSMT tasks. We therefore

combine the formality tagged FT data with the MT data without their tags (Figure 6.2b).

**MultiTask-random** investigates the impact of the similarity between formality transfer and bilingual examples. Selecting bilingual data which is similar to the GYAFC corpus is not necessarily beneficial for the FSMT task especially when French-English bilingual examples are drawn from a domain distant from the GYAFC corpus. In this configuration, we test how well our model performs FSMT if bilingual examples are randomly selected instead (Figure 6.2c).

## 6.2 Experimental Set-Up

**FT data:** We use the GYAFC corpus introduced by [Rao and Tetreault \(2018\)](#) as our FT data. This corpus consists of informal sentences from two domains of Yahoo Answers (i.e., *Entertainment and Music* (E&M) and *Family and Relationships* (F&R)) paired with their formal rewrites by humans. The train split consists of 105K informal-formal sentence pairs whereas the dev/test sets consist of roughly 10K/5K source-style sentences paired with four reference target-style human rewrites for both transfer directions.

**FSMT data:** We evaluate the FSMT models on a large-scale French to English (FR-EN) translation task. Examples are drawn from OpenSubtitles2016 ([Lison and Tiedemann, 2016](#)) which consists of movie and television subtitles and is thus more similar to the GYAFC corpus compared to news or parliament proceedings (e.g., MultiUN used by the reranking-based FSMT in Chapter 4). This is a noisy dataset

where aligned French and English sentences often do not have the same meaning, so we use a bilingual semantic similarity detector to select 20,005,000 least divergent examples from  $\sim 27.5$ M deduplicated sentence pairs in the original set (Vyas et al., 2018). Selected examples are then randomly split into a 20M training pool, a 2.5K dev set and a 2.5K test set.

**Preprocessing:** We apply the same pre-processing steps for bi-directional NMT systems (Chapter 5) to both FT and MT data: normalization, tokenization, truecasing, joint source-target BPE with 32,000 operations for NMT (Sennrich et al., 2016c), and sentence-filtering (length 50 cutoff) to parallel training data.

**NMT Configuration:** We use the standard attentional encoder-decoder architecture implemented in the `Sockeye` toolkit (Hieber et al., 2017) with the same settings introduced in Chapter 5. Our translation model uses a bi-directional encoder with a single LSTM layer (Bahdanau et al., 2015) of size 512, multilayer perceptron attention with a layer size of 512, and word representations of size 512. We apply layer normalization and tie the source and target embeddings as well as the output layer’s weight matrix. We add dropout to embeddings and RNNs of the encoder and decoder with probability 0.2. We train using the Adam optimizer with a batch size of 64 sentences and checkpoint the model every 1000 updates (Kingma and Ba, 2015). Training stops after 8 checkpoints without improvement of validation perplexity. We decode with a beam size of 5. We train four randomly seeded models for each experiment and combine them in a linear ensemble for decoding.

## 6.3 Evaluation Protocol

### 6.3.1 Automatic Evaluation

We evaluate both FT and FSMT tasks using BLEU (Papineni et al., 2002), which compares the model output with four reference target-style rewrites for FT and a single reference translation for FSMT. We select case-sensitive BLEU with standard WMT tokenization as our evaluation metric.<sup>2</sup> For FT, Rao and Tetreault (2018) show that BLEU correlates well with the overall system ranking assigned by humans. For FSMT, as explained earlier in Chapter 4, BLEU is an imperfect metric as it conflates mismatches due to translation errors and due to correct style variations. We therefore turn to human evaluation to isolate formality differences from translation quality.

### 6.3.2 Human Evaluation

Following the human evaluation protocol for the reranking-based FSMT (Chapter 4) and Rao and Tetreault (2018), we assess model outputs on three criteria: *formality*, *fluency* and *meaning preservation*. Since the goal of our evaluation is to compare models, our evaluation scheme asks workers to compare sentence pairs on these three criteria instead of rating each sentence in isolation. For FT, we compare the top performing NMT benchmark model in Rao and Tetreault (2018) with our best FT model. For FSMT, we compare outputs from three representative models:

---

<sup>2</sup>We use the script <https://github.com/EdinburghNLP/nematus/blob/master/data/multi-bleu-detok.perl>

NMT-constraint, MultiTask-random and PBMT-random.<sup>3</sup>

We collect human judgments using CrowdFlower.<sup>4</sup> Since we want native English speakers to perform this task, we restrict our set of annotators only to these three native English speaking countries: United States, United Kingdom, and Australia. We create a sample of 51 gold questions for each of the three criteria. Annotators have to continually maintain the accuracy of above 70% on these gold questions to be able to contribute to the task.

We collect judgments on 300 samples of each model output and we collect three judgments per sample (i.e., sentence pair). Given the three judgments per sample, we calculate the aggregate score using the weighted average:

$$\frac{\sum_{i=1}^3 \text{score}_i \times \text{trust}_i}{\sum_{i=1}^3 \text{trust}_i},$$

where  $\text{score}_i$  is the score given by an annotator and  $\text{trust}_i$  is our trust on that annotator. This trust is the accuracy of the annotator on the gold questions.

**Formality:** For FT, we want to measure the amount of style variation introduced by a model. Hence, we ask workers to compare the source-style sentence with its target-style model output. For FSMT, we want to measure the amount of style variation between two different translations by the same model. Hence, we ask workers to compare the “informal” English translation and the “formal” English translation of the same source sentence in French.<sup>5</sup>

---

<sup>3</sup>Note that we also compare with the English reference translation in Chapter 7.

<sup>4</sup><http://www.crowdfunder.com>

<sup>5</sup>Evaluating which systems produces the most (in)formal output is an independent question, and we will discuss it in Chapter 7.

Given two sentences, we ask workers to compare their formality using one of the following categories, regardless of fluency and meaning. We do not enumerate specific rules (e.g., typos or contractions) and encourage workers to use their own judgment.

Score	Category
2	Sentence 1 is much more formal than Sentence 2
1	Sentence 1 is more formal than Sentence 2
0	No difference or hard to say
-1	Sentence 2 is more formal than Sentence 1
-2	Sentence 2 is much more formal than Sentence 1

These categories are assigned scores in a symmetric range of [-2,2]. We randomly swap the two items in the pair so that annotators cannot guess which one is supposed to be more formal. When aggregating these scores, we recover the order, and sentence pairs with incorrect formality (e.g., the system’s informal output is actually annotated as more formal than its formal output) get negative scores.

**Fluency:** For both FT and FSMT tasks, we want to understand how fluent are the different model outputs. Hence, we ask workers to compare the fluency of two model outputs of the same target style. Similar to formality evaluation, we design a five point scale for comparing the fluency of two sentences, giving us a value between 0 and 2 for each sentence pair.

Given two sentences, we ask workers to compare their fluency using one of the following categories, regardless of style and meaning. We suggest that a sentence is fluent if it has a meaning and is coherent and grammatical well-formed. Fluency scores are aggregated in the same way as for formality scores.

Score	Category
2	Sentence 1 is much more fluent than Sentence 2
1	Sentence 1 is more fluent than Sentence 2
0	No difference or hard to say
-1	Sentence 2 is more fluent than Sentence 1
-2	Sentence 2 is much more fluent than Sentence 1

**Meaning Preservation:** For FT, we want to measure the amount of meaning preserved during formality transfer. Hence, we ask workers to compare the source-style sentence and the target-style model output. For FSMT, we want to measure the amount of meaning preserved between two different translations by the same model. Hence, we ask workers to compare the “informal” English translation and the “formal” English translation of the same source sentence in French. We design a four point scale to compare the meaning of two sentences ranging from the two being completely equivalent to the two being not equivalent, giving us a value between 0 and 3 for each sentence pair.

Given two sentences, we ask workers to answer “how much of the first sentence’s meaning is preserved in the second sentence”, regardless of style.

Score	Category
3	Equivalent since they convey the same key idea
2	Mostly equivalent since they convey the same key idea but differ in some unimportant details
1	Roughly equivalent since they share some ideas but differ in important details
0	Not equivalent since they convey different ideas

## 6.4 Formality Transfer Experiments

### 6.4.1 Baseline Models

We first compare baseline models from [Rao and Tetreault \(2018\)](#).

**PBMT** is a phrase-based machine translation model trained on the GYAFC corpus using a training regime consisting of self-training, data sub-selection and a large language model.

**NMT Baseline** uses OpenNMT-py ([Klein et al., 2017](#)). [Rao and Tetreault \(2018\)](#) use a pre-processing step to make source informal sentences more formal and source formal sentences more informal by rules such as re-casing. Word embeddings pre-trained on Yahoo Answers are also used.

**NMT Combined** is [Rao and Tetreault](#)'s best performing NMT model trained on the rule-processed GYAFC corpus, with additional forward and backward translations produced by the PBMT model.

### 6.4.2 Our Models

**NMT Baseline:** Our NMT baseline uses Sockeye instead of OpenNMT-py and is trained on raw datasets of two domains and two transfer directions.

**Bi-Directional FT:** Our initial bi-directional model is trained on bi-directional data from both domains with formality tags. It is incrementally augmented with three modifications to get the final multi-task model (i.e., MultiTask-tag-style as described in Section [6.1.3](#)): (1) We combine training sets of two GYAFC domains

Model	Informal→Formal		Formal→Informal	
	E&M	F&R	E&M	F&R
<i>Rao and Tetreault (2018)</i>				
PBMT	68.22	72.94	33.54	32.64
NMT Baseline	58.80	68.28	30.57	36.71
NMT Combined	68.41	74.22	33.56	35.03
<i>Ours</i>				
NMT Baseline	65.34	71.28	32.36	36.23
Bi-directional FT	66.30	71.97	34.00	36.33
+ training on E&M + F&R	69.20	73.52	35.44	37.72
+ ensemble decoding (×4)	71.36	74.49	36.18	38.34
+ multi-task learning (MultiTask-tag-style)	<b>72.13</b>	<b>75.37</b>	<b>38.04</b>	<b>39.09</b>

Table 6.1: Automatic evaluation of Formality Transfer with BLEU scores. The bi-directional model with three stacked improvements achieves the best overall performance. The improvement over the second best system is statistically significant at  $p < 0.05$  using bootstrap resampling (Koehn, 2004b).

(E&M+F&R) together and train a single model on the combination. (2) We use ensemble decoding by training four randomly seeded models on the combined data. (3) We add formality-tagged bilingual data and train the model using multi-task learning to jointly learn FT and FSMT. Suppose the amount of original bi-directional FT data is  $n$ , we always select  $kn$  bilingual data where  $k$  is an integer. We also upsample FT data to make it match the size of selected bilingual data.

### 6.4.3 Results

**Automatic Evaluation.** As shown in Table 6.1, our NMT baselines yield surprisingly better BLEU scores than those of Rao and Tetreault (2018) in most cases, even without using rule-processed source training data and pre-trained word embeddings. We attribute the difference to the more optimized NMT toolkit we use.

Initial bi-directional models outperform uni-directional models. This matches

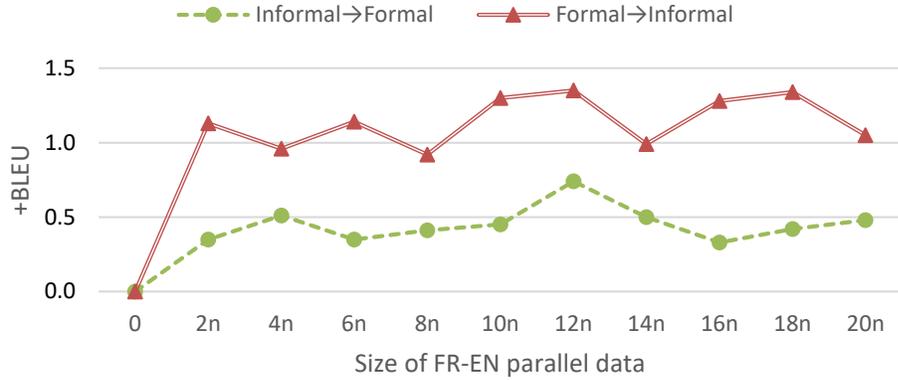
	Model A	Model B	Formality Diff Range = [-2,2]		Meaning Prsv. Range = [0,3]
			I→F	F→I	
FT	Source	NMT Combined	0.54	0.45	2.94
	Source	MultiTask-tag-style	0.59	<b>0.64</b>	2.92
FSMT	NMT-constraint I	NMT-constraint F	0.35		2.95
	MultiTask-random I	MultiTask-random F	0.32		2.90
	PBMT-random I	PBMT-random F	0.05		2.97

Table 6.2: Human evaluation of formality difference and meaning preservation. MultiTask-tag-style generates significantly more informal (F→I) English than NMT Combined ( $p < 0.05$  using the t-test, see Section 6.4.3). PBMT-random does not control formality effectively when comparing its informal (I) and formal (F) output (Section 6.5.2). Formality scores are relatively low because workers rarely choose “much more (in)formal”. All models preserve meaning equally well.

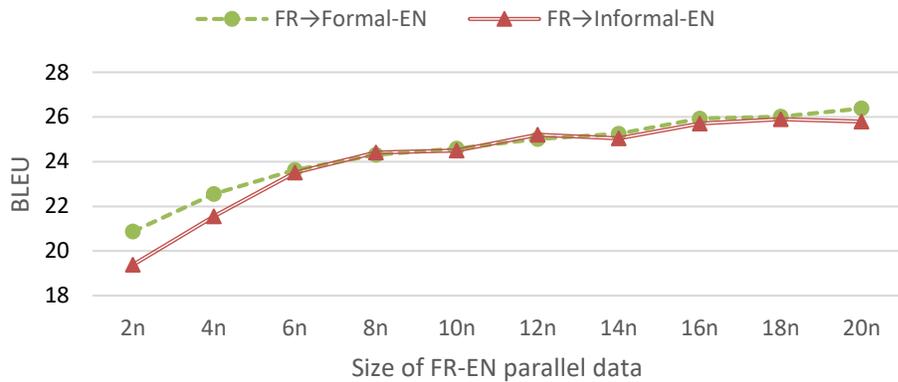
the behavior of bi-directional NMT in low-resource settings studied in Chapter 5 — we work with a relatively small amount of training data ( $\sim 50K$ ), and FT models benefit from doubling the size of training data without being confused by mixing two transfer directions. For the same reason, increasing the training data by combining two domains together improves performance further. Ensemble decoding is a consistently effective technique used by NMT and it enhances our NMT-based FT models as expected.

Incorporating the bilingual parallel data by multi-task learning yields further improvement. The target side of bilingual data is selected based on the closeness to the GYAFC corpus, so we hypothesize that the higher quality comes from better target language modeling by training on more English text.

**Human Evaluation.** The superior performance of the best FT model (i.e., MultiTask-tag-style) is also reflected in our human evaluation (see Table 6.2). It generates slightly more formal English (0.59 vs 0.54) and significantly more informal English



(a) BLEU improvements for formality transfer.



(b) BLEU scores for machine translation.

Figure 6.3: BLEU improvements or scores for four transfer/translation directions vs. the size of FR-EN parallel data.  $n$  in x-axis equals to the original size of bi-directional style transfer training data. Formality transfer improves with bilingual data and the performance reaches the plateau quickly. The translation quality increases monotonically with the size of training data.

(0.64 vs 0.45) than NMT Combined. This is consistent with BLEU differences in Table 6.1 which show that MultiTask-tag-style yields bigger improvements when transferring formal language to informal. Both models have good quality with respect to meaning preservation (2.94 vs 2.92) and workers can hardly find any fluency difference between outputs of these two models by assigning 0.03 in average in the fluency test (0 means no difference).

**Impact of Bilingual Data Size.** We evaluate the impact of selected bilingual data size on the combination of development sets from two domains in GYAFC and show the results in Figure 6.3. The quality of formality transfer improves instantly when using bilingual data and it soon reaches the plateau when more data is used. Meanwhile, the translation quality increases monotonically with the size of training data. The optimal point is a hyper-parameter that can be determined on the development set. We empirically choose  $n = 12$  since it works best for formality transfer and yields reasonable translation quality.

#### 6.4.4 Qualitative Analysis

We manually inspect 100 randomly selected samples from our evaluation set and compare the target-style output of our best model (MultiTask-tag-style) with that of the best baseline model (NMT-Combined) from Rao and Tetreault (2018). Table 6.3 shows some samples representative of the trends we find for `informal`→`formal` (6.3a) and `formal`→`informal` (6.3b) tasks.

In majority of the cases, the two models produce similar outputs as can be expected since they use similar NMT architectures. In cases where the two outputs differ, in the I→F task, we find that our model produces a more formal output by introducing phrasal level changes (first sample in Table 6.3a) or by moving phrases around (second sample in Table 6.3a), both of which happen frequently during machine translation, thus showcasing the benefit of our multi-task approach. Our model very often makes the output sentence more complete (and thereby more

<b>6.3a: informal <math>\rightarrow</math> formal</b>	
Original I	chill out sweetie everything will be just fine eventually
1 NMT-Combined F	Can you chill out sweetie everything will be just fine eventually.
MultiTask-tag-style F	Calm down, sweetie, everything will be fine eventually.
Original I	Dakota Fanning....I know that she is only 12 but she is really famous.
2 NMT-Combined F	Dakota Fanning,i know that she is only twelve, but she is famous.
MultiTask-tag-style F	I know that Dakota Fanning is only twelve, but she is really famous.
Original I	depends....usully they are about ur personalty but not wat ur gonna do iwith ur life.
3 NMT-Combined F	Depends.usully they are about your personalty, but not what your going to do iwith your life.
MultiTask-tag-style F	It depends. They are about your personalty, but not what you are going to do with your life.
Original I	THAT DEPENDS...ARE YOU A HOTTIE W/A BODY?
4 NMT-Combined F	That depends, are you a hottie with a body?
MultiTask-tag-style F	That depends. Are you a HOTTIE W / A BODY?
<b>6.3b: formal <math>\rightarrow</math> informal</b>	
Original F	Therefore I would say that they do succeed but not frequently. I hope this is helpful.
1 NMT-Combined I	So I would say that they do failing but not frequently, I hope this is helps.
MultiTask-tag-style I	so i would say they do it but not all the time, hope this helps.
Original F	I am simply inquiring because people behave as though they are no longer interested in them.
2 NMT-Combined I	I am just asking because people act as though they are no longer interested in them.
MultiTask-tag-style I	I'm just asking because people act like they don't like them anymore.
Original F	Hello, I am interested in visiting your country.
3 NMT-Combined I	Hi, I'm interested in visiting your country.
MultiTask-tag-style I	hi, I'm going to go to your country.

Table 6.3: Sample model outputs for the Formality Transfer (FT) task.

formal) by inserting pronouns like ‘it’ or ‘they’ at the start of the sentence or by removing conjunctions like ‘usually’, ‘and’, ‘but’, ‘however’ from the beginning of a sentence (sample three in Table 6.3a). Likewise, in the  $F \rightarrow I$  task, our model produces more informal sentences compared to the baseline by introducing more phrasal level changes (first and second sample in Table 6.3b).

**Error analysis:** In the  $I \rightarrow F$  task, our model performs worse than the baseline when the original informal sentence consists of all uppercased words (fourth sample in Table 6.3a). This is primarily because the baseline model pre-lowercases them using rules, whereas, we rely on the model to learn this transformation, and it fails to do so for less frequent words. In the  $F \rightarrow I$  task, in trying to produce more informal outputs, our model sometimes fails to preserve the original meaning of the sentence (third sample in Table 6.3b). In both tasks, very often our model fails to make transformations for some pairs like (‘girls’, ‘women’), which the baseline model is very good at. We hypothesize that this could be because for these pairs, human rewriters do not always agree on one of the words in the pair being more informal/formal. This makes our model more conservative in making changes because our bi-directional model combines FT data from both directions and when the original data contains instances where these words are not changed, we double that and learn to copy the word more often than change it.

## 6.5 Formality-Sensitive Machine Translation Experiments

### 6.5.1 Models

**NMT-constraint:** We first evaluate the standard NMT model with side constraints introduced in Section 6.1.2 and then compare it with three variants of FSMT models using multi-task learning as described in Section 6.1.3 (i.e., **MultiTask-tag-style**, **MultiTask-style** and **MultiTask-random**). The best performing system for FT is MultiTask-tag-style with  $12n$  ( $\sim 2.5M$ ) bilingual pairs. For fair comparison, we select this size of bilingual data for all FSMT models either by data selection or randomly.

**PBMT-random:** We also compare these models with the PBMT-based FSMT system described in Chapter 4. Instead of tagging sentences in a binary fashion, this system scores each sentence using a lexical formality model. It requests a desired formality score for translation output and re-ranks  $n$ -best translation hypotheses by their closeness to the desired formality level. We adapt this system to our evaluation scenario — we calculate median scores for informal and formal data (i.e.,  $-0.41$  and  $-0.27$  respectively) in GYAFC respectively by a PCA-LSA-based formality model and use them as desired formality levels.<sup>6</sup> The bilingual training data is randomly selected.

---

<sup>6</sup>The PCA-LSA-based formality model achieves lowest root-mean-square error on a scoring task of sentential formality as listed on <https://github.com/xingniu/computational-stylistic-variations>.

Model	+Tag?	Random?	FR→Formal-EN	FR→Informal-EN
NMT-constraint	✓		27.15	26.70
MultiTask-tag-style	✓		25.02	25.20
MultiTask-style			23.25	23.41
MultiTask-random		✓	25.24	25.14
PBMT-random		✓	29.12	29.02

Table 6.4: BLEU scores of various FSMT models. “+Tag” indicates using formality tags for bilingual data. “Random” indicates using randomly selected bilingual data.

## 6.5.2 Results

**Automatic Evaluation.** We compute BLEU scores on the FSMT test set for all models as a sanity check on translation quality. Because there is only one reference translation of unknown style for each input sentence, these BLEU scores conflate translation errors and stylistic mismatch, and are therefore not sufficient to evaluate FSMT performance. We include them for completeness here, as indicators of general translation quality, and will rely on human evaluation as primary evaluation method. As can be seen in Table 6.4, changing the formality level for a given system yields only small differences in BLEU. We select MultiTask-random as the representative of multi-task FSMT since it achieves competitive BLEU scores among multi-task models and contains more in-domain translation data. We compare MultiTask-random with NMT-constraint and PBMT-random during our human evaluation.

**Human Evaluation.** Table 6.2 shows that neural models control formality significantly better than PBMT-random (0.35/0.32 vs. 0.05). They also introduce more changes in translation: with NMT models,  $\sim 80\%$  of outputs change when only the input formality changes, while that is only the case for  $\sim 30\%$  of outputs with PBMT-

random. Among neural models, MultiTask-random and NMT-constraint have similar quality in controlling the output formality (0.32 vs. 0.35) and preserving meaning (2.90 vs. 2.95). They are also equally fluent as judged by humans. Interestingly, multi-task learning helps MultiTask-random perform similarly as NMT-constraint with simpler examples that do not require the additional step of data selection to generate formality tags.

### 6.5.3 Qualitative Analysis

We randomly sample 100 examples from our test set and manually compare the formal and the informal translations of the French source by MultiTask-random, NMT-constraint and PBMT-random. Table 6.5 shows representative examples of the observed trends.

We find that in most cases, the difference between the formal and informal style translations is very minor in PBMT-random model, better in NMT-constraint model and the best in our MultiTask-random model (first sample in the table). In general, our MultiTask-random model does a good job of making very large changes while transferring the style, especially into informal (second sample in the table). We hypothesize that this is because our joint model is trained on the GYAFC corpus which consists of parallel sentences that differ heavily in style.

**Error analysis:** All FSMT models perform well in terms of meaning preservation, yet the human scores are not perfect (Table 6.2). They occasionally change not only the style but also the meaning of the input (e.g., the third sample of MultiTask-

1	French Source	Impossible d’avoir accès à internet ici.
	English Reference	I don’t know if you’ve tried yet, but it’s impossible to get on the internet up here.
	MultiTask-random Formal	It is impossible to have access to the internet here.
	NMT-constraint Formal	It’s impossible to have access to the Internet here.
	PBMT-random Formal	I can’t access to the internet here.
	MultiTask-random Informal	Impossible to get to the internet here.
	NMT-constraint Informal	Couldn’t have accessed the internet here.
	PBMT-random Informal	I can’t access to the internet here.
2	French Source	Abstenez-vous de tout commentaire et répondez à la question, chef Toohey.
	English Reference	Refrain from the commentary and respond to the question, Chief Toohey.
	MultiTask-random Formal	You need to be quiet and answer the question, Chief Toohey.
	NMT-constraint Formal	Please refrain from any comment and answer the question, Chief Toohey.
	PBMT-random Formal	Please refrain from comment and just answer the question, the Tooheys’s boss.
	MultiTask-random Informal	Shut up and answer the question, Chief Toohey.
	NMT-constraint Informal	Please refrain from comment and answer the question, chief Toohey.
	PBMT-random Informal	Please refrain from comment and answer my question, Tooheys’s boss.
3	French Source	Essaie de présenter des requêtes supplémentaires dès que tu peux.
	English Reference	Try to file any additional motions as soon as you can.
	MultiTask-random Formal	You should try to introduce the sharks as soon as you can.
	NMT-constraint Formal	Try to present additional requests as soon as you can.
	PBMT-random Formal	Try to introduce any additional requests as soon as you can.
	MultiTask-random Informal	Try to introduce sharks as soon as you can.
	NMT-constraint Informal	Try to introduce extra requests as soon as you can.
	PBMT-random Informal	Try to introduce any additional requests as soon as you can.

Table 6.5: Sample model outputs for the Formality-Sensitive Machine Translation (FSMT) task.

random in Table 6.5). This motivates future work that penalizes meaning changes more explicitly during training. In general, none of the models do a good job of changing the style when the source sentence is not skewed in one style. For example, consider the French sentence “Combien de fois vous l’ai-je dit?” and its English reference translation “How many times have I told you, right?”. All models produce the same translation “How many times did I tell you?”. In such cases, changing style requires heavier editing or paraphrasing of the source sentence that our current models are unable to produce.

## 6.6 Summary

We explored the use of multi-task learning to perform monolingual FT and bilingual FSMT jointly. Using French-English translation and English style transfer data, we showed that the joint model is able to learn from both style transfer parallel examples and translation parallel examples. On the FT task, the joint model significantly improved the quality of transfer between formal and informal styles in both directions, compared to prior work (Rao and Tetreault, 2018). This also represents a strong baseline for follow up work in formality transfer: other results show that using more sophisticated approaches, such as post-editing (Ge et al., 2019) and constrained decoding (Kajiwara, 2019), do not help as much as the parallel data introduced via multi-task learning. The joint model interestingly also learned to perform FSMT without being explicitly trained on style-annotated translation examples. On the FSMT task, our model outperformed previously proposed PBMT

model and performed on par with a neural model with side-constraints, which requires more involved data selection. However, neural FSMT models sometimes produced translations disobeying the source meaning, and the formality-control intensity could further be improved.

## Chapter 7: Neural FSMT with Synthetic Supervision

Building an FSMT system ideally requires training triplets consisting of a bilingual sentence pair labeled with target language formality. However, bilingual parallel corpora do not come with formality annotations, and parallel corpora of a given provenance do not have a uniform style. The multi-task FSMT models introduced in Chapter 6 are presented with samples where one element of the triplet is always missing. Therefore, it is trained to perform FSMT in a zero-shot fashion, and it sometimes produces translations that are inappropriate for the desired formality, or that match the formality level but do not preserve the source meaning.

We hypothesize that exposing multi-task models to complete training triplets should improve the quality of formality-sensitive language generation, so that formal and informal outputs differ from each other more and formality rewrites do not introduce translation errors. To this end, we introduce an approach to predict the target formality for a given parallel sentence pair. This approach simulates direct supervision on the fly for end-to-end training. We also explore the possibility of generating a synthetic ground truth translation given an input language sentence and the desired formality and present a variant of side constraints ([Sennrich et al.](#),

2016a) that improves formality control.<sup>1</sup>

We conduct a comprehensive automatic and human evaluation of the resulting FSMT systems: (1) We measure translation quality and quantify differences between translations at opposite formality using automatic metrics. (2) We rely on human judgments to assess the performance of meaning preservation and formality control between a strong multi-task baseline and the most promising of the proposed models. (3) We analyze outputs qualitatively to illustrate how formality is marked in model outputs. Results show that our best model trained with synthetic supervision outperforms prior neural FSMT models. It produces translations that better match desired formality levels while preserving the source meaning.

## 7.1 Approach

Recall that FSMT requires producing the most likely translation at the given formality level  $\ell$ :

$$\hat{\mathbf{Y}} = \arg \max_{\mathbf{Y}_\ell} P(\mathbf{Y}_\ell | \mathbf{X}, \ell; \boldsymbol{\theta}). \quad (7.1)$$

Ideally, the FSMT model should be trained on triplets  $(\mathbf{X}, \ell, \mathbf{Y}_\ell)_{1\dots N}$ , but in practice, such training data is not easy to acquire. In Chapter 6, we tackle this problem by training a cross-lingual machine translation model (French $\rightarrow$ English) and a monolingual bidirectional formality transfer model (Formal-English $\leftrightarrow$ Informal-English) jointly. Specifically, the model is trained on the combination of  $(\mathbf{X}, \mathbf{Y})_{1\dots N_1}$  and  $(\mathbf{Y}_{\bar{\ell}}, \ell, \mathbf{Y}_\ell)_{1\dots N_2}$ , where  $\mathbf{Y}_{\bar{\ell}}$  and  $\mathbf{Y}_\ell$  have opposite formality levels. The joint model is

---

<sup>1</sup>Data and code are available at <https://github.com/xingniu/multitask-ft-fsmt>.

able to perform zero-shot FSMT by optimizing  $\mathcal{L}_{MT} + \mathcal{L}_{FT}$ , where

$$\mathcal{L}_{MT} = \sum_{(\mathbf{x}, \mathbf{y})} \log P(\mathbf{Y}|\mathbf{X}; \boldsymbol{\theta}), \quad (7.2)$$

$$\mathcal{L}_{FT} = \sum_{(\mathbf{Y}_{\bar{\ell}}, \ell, \mathbf{Y}_{\ell})} \log P(\mathbf{Y}_{\ell} | \mathbf{Y}_{\bar{\ell}}, \ell; \boldsymbol{\theta}). \quad (7.3)$$

### 7.1.1 Controlling the Output Language Formality

FSMT shares the goal of producing output sentences of a given formality with monolingual formality style transfer tasks. In both cases, the source sentence usually carries its own style and the model should be able to override it with the independent style  $\ell$ . This is achieved by using an attentional sequence-to-sequence model with side constraints (Sennrich et al., 2016a), i.e., attaching a style tag (e.g., `<2Formal>`) to the beginning of each source example. Here, we attach style tags to both source and target sequences.

Sennrich et al. (2016a) hypothesize that source-side tags control the target style because the model “learns to pay attention to the side constraints”, but it has not been verified empirically. We hypothesize that the source style tag also influences the encoder hidden states, and providing a target-side tag lets the decoder benefit from encoding style more directly. This method is analogous to replacing begin-of-sequence (`<BOS>`) embeddings in the target with style embeddings (Lample et al., 2019), but requires zero-modification to the architecture: the model easily learns to predict the target tag by training on tagged data with a standard architecture.

## 7.1.2 Synthetic Supervision

Prior work on multilingual NMT shows that the translation quality on zero-shot tasks often significantly lags behind when supervision is provided (Johnson et al., 2017). We address this problem by simulating the supervision, i.e., generating synthetic training triplets  $(\mathbf{X}, \ell, \mathbf{Y})$  by using the FSMT model itself to predict the missing element of the triplet from parallel sentence pairs  $(\mathbf{X}, \mathbf{Y})$ . We introduce two novel approaches to generate synthetic triplets, namely Online Style Inference and Online Target Inference.

### 7.1.2.1 Online Style Inference (OSI)

Given a translation example  $(\mathbf{X}, \mathbf{Y})$ , we view predicting the formality of  $\mathbf{Y}$ , i.e.,  $\ell_{\mathbf{Y}}$ , as unsupervised classification using only the pre-trained FSMT model.

As illustrated in Figure 7.1, we use FSMT to produce both informal and formal translations of the same input,  $\mathbf{Y}_{\text{I}} = \text{FSMT}(\mathbf{X}, \ell_{\text{I}})$  and  $\mathbf{Y}_{\text{F}} = \text{FSMT}(\mathbf{X}, \ell_{\text{F}})$  respectively.<sup>2</sup> We hypothesize that the style of the reference translation  $\mathbf{Y}$  can be predicted based on its distance from these two translations. For example, if  $\mathbf{Y}$  is formal, it should be closer to  $\mathbf{Y}_{\text{F}}$  than  $\mathbf{Y}_{\text{I}}$ . We measure the closeness by cross-entropy difference (CED, Moore and Lewis, 2010), i.e., we calculate the difference of their per-token cross-entropy scores,  $\text{CED}(\mathbf{Y}_{\text{I}}, \mathbf{Y}_{\text{F}}) = H_{\mathbf{Y}}(\mathbf{Y}_{\text{I}}) - H_{\mathbf{Y}}(\mathbf{Y}_{\text{F}})$ . The larger it is, the closer  $\mathbf{Y}$  is to  $\mathbf{Y}_{\text{F}}$ .

---

<sup>2</sup> $\mathbf{Y}_{\text{I}}$  and  $\mathbf{Y}_{\text{F}}$  are generated with the *teacher forcing* strategy (Williams and Zipser, 1989) given the ground-truth  $\mathbf{Y}$ .

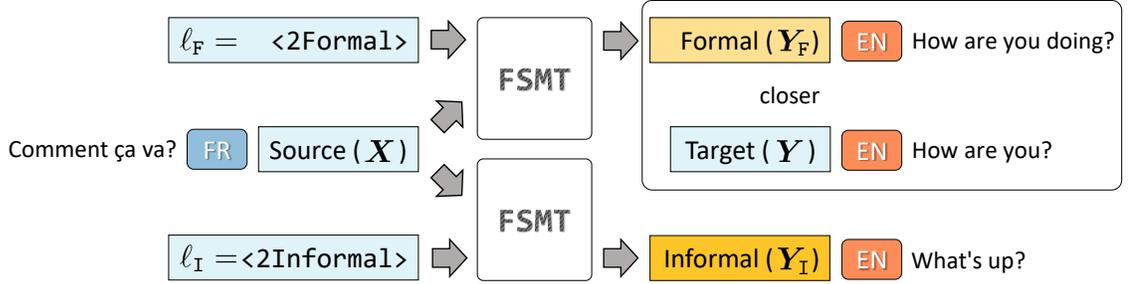


Figure 7.1: Online Style Inference. Given a translation example  $(\mathbf{X}, \mathbf{Y})$ , FSMT produces both informal and formal translations of  $\mathbf{X}$ , i.e.,  $\mathbf{Y}_I = \text{FSMT}(\mathbf{X}, \ell_I)$  and  $\mathbf{Y}_F = \text{FSMT}(\mathbf{X}, \ell_F)$ .  $\mathbf{Y}$  is labeled as formal since it is closer to  $\mathbf{Y}_F$  than  $\mathbf{Y}_I$ .

Given a positive threshold  $\tau$ , we label  $\ell_Y = \langle 2\text{Informal} \rangle$  if  $\text{CED}(\mathbf{Y}_I, \mathbf{Y}_F) < -\tau$ , label  $\ell_Y = \langle 2\text{Formal} \rangle$  if  $\text{CED}(\mathbf{Y}_I, \mathbf{Y}_F) > \tau$ , and label  $\ell_Y = \langle 2\text{Unknown} \rangle$  otherwise. The threshold  $\tau$  is chosen dynamically for each mini-batch, and it is equal to the mean of absolute token-level CED of all tokens within a mini-batch. Finally, we are able to generate a synthetic training sample,  $(\mathbf{X}, \ell_Y, \mathbf{Y})$ , on the fly and optimize  $\mathcal{L}_{FT} + \mathcal{L}_{OSI}$ , where

$$\mathcal{L}_{OSI} = \sum_{(\mathbf{X}, \ell_Y, \mathbf{Y})} \log P(\mathbf{Y} | \mathbf{X}, \ell_Y; \boldsymbol{\theta}). \quad (7.4)$$

### 7.1.2.2 Online Target Inference (OTI)

Given the bilingual parallel sentence pair  $(\mathbf{X}, \mathbf{Y})$  and a randomly selected target formality  $\ell$  from  $\{\langle 2\text{Informal} \rangle, \langle 2\text{Formal} \rangle\}$ , we can use the FSMT model to produce a formality-constrained translation  $\mathbf{Y}_\ell^1 = \text{FSMT}(\mathbf{X}, \ell)$ . We exploit the multi-task nature of the FSMT model to estimate the quality of  $\mathbf{Y}_\ell^1$  indirectly without supervision: the FSMT model can also manipulate the formality level of the target side  $\mathbf{Y}$  via monolingual formality transfer to produce  $\mathbf{Y}_\ell^2 = \text{FT}(\mathbf{Y}, \ell)$ . We hypothesize that the predictions made by these two different paths should be

consistent.

The quality of  $\mathbf{Y}_\ell^2$  is presumably more reliable than  $\mathbf{Y}_\ell^1$ , because the transfer model (which is embedded in the joint model) is trained with direct supervision. We empirically get  $\mathbf{Y}_\ell^2$  via greedy search on the fly during the training and use it as the label. Finally, we optimize  $\mathcal{L}_{MT} + \mathcal{L}_{FT} + \alpha\mathcal{L}_{OTI}$ , where

$$\mathcal{L}_{OTI} = \sum_{(\mathbf{x}, \ell, \mathbf{Y}_\ell^2)} \log P(\mathbf{Y}_\ell^2 | \mathbf{x}, \ell; \boldsymbol{\theta}). \quad (7.5)$$

Online Target Inference is a harder task than Online Style Inference since it requires generating language as opposed to making a formality prediction.

## 7.2 Auxiliary English Formality Control Evaluation

Before investigating how to improve FSMT with synthetic supervision, we investigate whether alternatives to side constraints would be beneficial to formality control. Our goal is to determine a solid approach for formality control before adding synthetic supervision. For simplicity, we conduct this auxiliary evaluation of formality control on the monolingual style transfer task.

**Task** Our task aims to test systems ability to produce a formal or an informal paraphrase for a given English sentence of arbitrary style. It is derived from formality transfer (Rao and Tetreault, 2018), where models transfer sentences from informal to formal ( $\mathbf{I} \rightarrow \mathbf{F}$ ) or from formal to informal ( $\mathbf{F} \rightarrow \mathbf{I}$ ). The tests of flipping the formality levels, as we have done so far in Chapter 6, only evaluate a model’s ability in learning mappings between informal and formal languages. We addition-

ally evaluate the ability of systems to preserve formality on informal to informal ( $I \rightarrow I$ ) and formal to formal ( $F \rightarrow F$ ) tasks. This four-way formality rewriting setting is **particularly relevant to the FSMT task**, where the style of the source sentence is arbitrary.

We also use the GYAFC corpus (Rao and Tetreault, 2018) as in Chapter 6 for this evaluation. This corpus consists of informal sentences from Yahoo Answers paired with their formal rewrites by humans. The train split consists of 105K informal-formal sentence pairs whereas the dev/test sets consist of roughly 10K/5K source-style sentences paired with four reference target-style human rewrites for both transfer directions, i.e.,  $I \rightarrow F$  and  $F \rightarrow I$ . For formality preserving tasks, the output is compared with the input sentence in the test set.

**Models** All models are trained on bidirectional data, which is constructed by swapping the informal and formal sentences of the parallel GYAFC corpus and appending the swapped version to the original (the model configuration and training set-up are exactly the same as for the FSMT experiments and will be described in detail in Section 7.3.3). The formality of each target sentence represents the desired input style.

We first implement a baseline method which is trained only on the bidirectional data without showing the target formality (denoted as None). Next, we conduct an ablation study on the side constraint method to examine the hypothesis that model learns to pay attention to the tags by comparing TAG-SRC, TAG-SRC-BLOCK, and our proposed variant TAG-SRC-TGT. TAG-SRC is the standard method that

Model	Formality Transfer			
	I→F		F→I	
None	70.63 ± 0.23		37.00 ± 0.18	
TAG-SRC	72.16 ± 0.34	Δ	37.67 ± 0.11	Δ
TAG-SRC-BLOCK	72.00 ± 0.05	-0.16	37.38 ± 0.12	-0.29
TAG-SRC-TGT	72.29 ± 0.23	+0.13	37.62 ± 0.37	-0.05
	Formality Preservation			
	I→I		F→F	
None	54.54 ± 0.44		58.98 ± 0.93	
TAG-SRC	66.87 ± 0.58	Δ	78.78 ± 0.37	Δ
TAG-SRC-BLOCK	65.46 ± 0.29	<b>-1.41</b>	76.72 ± 0.39	<b>-2.06</b>
TAG-SRC-TGT	67.81 ± 0.41	<b>+0.94</b>	79.34 ± 0.55	<b>+0.56</b>

Table 7.1: BLEU scores for variants of side constraint in controlling style on all formality transfer and preservation directions. We report mean and standard deviation over five randomly seeded models.  $\Delta$ BLEU between each model and the widely used TAG-SRC methods show that (1) blocking the visibility of source tags from the encoder (TAG-SRC-BLOCK) limits its formality control ability; (2) using style tags on both source and target sides (TAG-SRC-TGT) helps control formality better, especially for formality preservation tasks.

attaches tags to the source, while TAG-SRC-BLOCK blocks the visibility of the tag embeddings from the encoder but retains their connections to the decoder via the attention mechanism. TAG-SRC-TGT attaches tags to both sides. We train five randomly seeded models for each method and report the mean and standard deviation of test BLEU (Table 7.1).

**Results** Comparing with methods acknowledging the target formality (i.e., TAG-SRC\*), the baseline method gets slightly lower BLEU scores when it learns to flip the formality on I→F and F→I tasks. However, it performs much worse (10-20 BLEU points lower) on I→I and F→F tasks since flipping the formality is harmful.

TAG-SRC-BLOCK lags behind TAG-SRC, especially for formality preservation

tasks (1-2 BLEU points lower). This discrepancy indicates that the attention mechanism only contributes a portion of the control ability. On the other hand, our proposed variant TAG-SRC-TGT performs better than TAG-SRC on 3/4 tasks (i.e.,  $I \rightarrow F$ ,  $I \rightarrow I$ , and  $F \rightarrow F$ ).

Taken together, these observations show that the impact of tags is not limited to the attention model, and their embeddings influence the hidden representations of encoders and decoders positively. The auxiliary evaluation thus confirms that adding style tags to both source and target sequences is a good approach to model monolingual formality transfer, and therefore motivates using it in our FSMT models as well.

### 7.3 FSMT Evaluation Set-Up

As mentioned in Chapter 4 and 6, evaluating FSMT systems requires evaluating whether their outputs correctly convey the meaning of the source, and whether the differences between their formal and informal outputs are indicative of formality. Since translations of the same text into formal and informal versions are not readily available, we use single reference translation of source sentences from diverse provenances to automatically evaluate the translation quality and output diversity of our systems. This **automatic evaluation** is imperfect as comparing against a single reference translation of arbitrary style does not let us separate translation errors from correct formal or informal paraphrases. We use the automatic evaluation during system development and to select a subset of models for **manual evaluation**.

We enhance the evaluation protocol in Chapter 6 by providing more accurate and explicit assessments.

### 7.3.1 Tasks and Data

**Test Sets** We still evaluate FSMT approaches on the French-English translation task as in previous chapters, but we choose two standard test sets for their higher quality than the held-out data from noisy training corpora. WMT newstest2014<sup>3</sup> and MSLT conversation test set<sup>4</sup> we use capture both formal and informal language. Each test set contains different formality levels, but the written language used in news stories is typically more formal than the spoken language used in conversations. As a result, for the first pass automatic evaluation, we assume that reference translations from newstest2014 are overall more formal, while references from MSLT are overall more informal. Human evaluation is then conducted without making this assumption.

**Training Sets** Following Chapter 6, we use OpenSubtitles2016 (Lison and Tiedemann, 2016), which consists of movie and television subtitles, covers a wider spectrum of styles, but overall tends to be informal since it primarily contains conversations. Again, we use a bilingual semantic similarity detector to select 16M least divergent examples from  $\sim 27.5$ M deduplicated sentence pairs in the original set (Vyas et al., 2018).<sup>5</sup> Since we focus on FSMT in this chapter, we continue the

---

<sup>3</sup><http://www.statmt.org/wmt14/test-full.tgz>

<sup>4</sup><https://www.microsoft.com/en-us/download/details.aspx?id=54689>

<sup>5</sup>We select slightly less examples (16M here vs. 20M in Chapter 6) but use all 16M examples for training.

Corpus	# sentences	# EN tokens
OpenSubtitles2016	16,000,000	171,034,255
Europarl.v7	1,670,324	39,789,959
News-Commentary.v14	276,358	6,386,435
WMT newstest2014	3,003	72,435
MSLT test	3,543	31,338

Table 7.2: Statistics of French-English corpora.

routine of building a reranking-based FSMT system (Chapter 4) and train models on two more parallel corpora with diverse formality: (1) Europarl.v7 (Koehn, 2005), which is extracted from the proceedings of the European Parliament, and tends to be more formal text; (2) News-Commentary.v14 (Bojar et al., 2018). The GYAFC corpus is also used to train multi-task models.

**Preprocessing** We apply the same pre-processing steps for the multi-task FSMT models (Chapter 6) here: normalization, tokenization, true-casing, joint source-target BPE with 50,000 operations (Sennrich et al., 2016c) and sentence-filtering (length 50 cutoff) to parallel training data.<sup>6</sup> Itemized data statistics after preprocessing can be found in Table 7.2. The MSLT data is pre-processed by removing duplicated and ill-encoded sentences.<sup>7</sup>

### 7.3.2 Baseline Models

We start with building two MT-only baseline models. The first is a standard **NMT** model which is trained with non-tagged French-English parallel data. The

<sup>6</sup>More BPE operations are used in this chapter because a larger vocabulary yields better high-resource performance.

<sup>7</sup>27% of the sentence pairs are duplicated, and the pre-processing script is released along with the source code.

second is **NMT DS-Tag** introduced in Chapter 6. It performs data selection on French-English training examples  $(\mathbf{X}, \mathbf{Y})$  using CED in a standard way: it pre-trains language models for informal and formal English in the formality transfer training data and calculates  $\text{CED}(\mathbf{Y}) = H_{\text{informal}}(\mathbf{Y}) - H_{\text{formal}}(\mathbf{Y})$ . Since we aim at using all parallel data, for fair comparison, we also conduct three-way tagging as introduced in Section 7.1.2.1. An NMT model is then trained with the formality-tagged training pairs.

Next, we use the multi-task FSMT models in Chapter 6 as stronger baselines.<sup>8</sup> The first version is **Multi-Task**. It performs zero-shot FSMT by training translation and formality transfer jointly. The second is **Multi-Task DS-Tag**, which is the combination of Multi-Task and NMT DS-Tag and is trained on both tagged translation pairs and formality transfer pairs. This method is similar to Online Style Inference in terms of tagging training examples using CED. However, Multi-Task DS-Tag uses standard offline language models while Online Style Inference can be interpreted as using source-conditioned online language models.

### 7.3.3 Implementation Details

We build NMT models upon the attentional RNN encoder-decoder architecture (Bahdanau et al., 2015) implemented in the Sockeye toolkit (Hieber et al., 2017) with the same settings introduced in Chapter 6. Our translation model uses a bi-directional encoder with a single LSTM layer of size 512, multilayer perceptron

---

<sup>8</sup>We considered a pivoting approach (i.e., machine translation followed by formality transfer) in preliminary experiments, but it consistently underperforms multi-task baselines.

attention with a layer size of 512, and word representations of size 512. We apply layer normalization (Ba et al., 2016), add dropout to embeddings and RNNs (Gal and Ghahramani, 2016) with probability 0.2, and tie the source and target embeddings as well as the output layer’s weight matrix (Press and Wolf, 2017). We train using the Adam optimizer (Kingma and Ba, 2015) with a batch size of 64 sentences and we checkpoint the model every 1000 updates. The learning rate for baseline models is initialized to 0.001 and reduced by 30% after 4 checkpoints without improvement of perplexity on the development set. Training stops after 10 checkpoints without improvement.

We build our models by fine-tuning Multi-Task with the dedicated synthetically supervised objectives described in Section 7.1.2, inheriting all settings except the learning rate which is re-initialized to 0.0001.

## 7.4 Automatic Evaluation of FSMT

### 7.4.1 Lessons from BLEU

We evaluate our systems by producing formal and informal translations for both the WMT and MSLT test sets, and we compare outputs against the single reference translation using BLEU.<sup>9</sup> As explained earlier in Chapter 4 and 6, this is an incomplete evaluation of FSMT, but it nevertheless provides simple sanity checks during system development: (1) Do BLEU scores of FSMT models remain close to that of formality-agnostic baselines, indicating that translation quality is roughly

---

<sup>9</sup>We report case-insensitive BLEU with the WMT standard ‘13a’ tokenization using SacreBLEU (Post, 2018).

maintained? (2) Do FSMT models obtain higher BLEU scores for formal outputs on WMT (where we expect references to be more formal) and higher BLEU scores for informal outputs on MSLT (where we expect references to be more informal)?

In Table 7.3, we first compare BLEU scores horizontally for each model. All FSMT systems achieve better scores when the formality level given as input to the system matches the nature of the text being translated. For example, formal translations are better for WMT news while informal translations are better for MSLT conversations.

$\Delta$ BLEU scores between informal and formal outputs show that multi-task models generate more dissimilar translations. However,  $\Delta$ BLEU does not show consistent trends across techniques and test sets, because it is a roundabout evidence of the sequence dissimilarity: it uses the reference as a proxy. We therefore quantify the differences between formal and informal outputs for each system more directly in Section 7.4.2.

Next, we compare BLEU scores vertically among models. Our proposed systems get relatively lower scores than baselines, which indicates that their outputs are more different from the reference translations.

## 7.4.2 Quantifying Differences Between Formal and Informal Outputs

**Metrics** We introduce the **Lexical and Positional Differences** (LEPOD) score to quantify the surface differences between the formal and informal outputs of a given system.

	WMT					MSLT				
	Informal	Formal	$\Delta$		PoD	Informal	Formal	$\Delta$		PoD
	BLEU	BLEU	BLEU	LED		BLEU	BLEU	BLEU	LED	
<i>MT-only Baselines</i>										
NMT	28.63	28.63	0	0	0	47.83	47.83	0	0	0
NMT DS-Tag	28.24	28.95	0.71	9.27	6.44	47.60	47.24	0.36	8.18	1.10
<i>Multi-task Baselines</i>										
Multi-Task	27.75	28.39	0.64	10.89	7.76	47.55	45.08	2.47	11.97	1.41
Multi-Task DS-Tag	27.65	29.12	1.47	11.51	8.35	47.46	46.66	0.80	10.29	1.54
<i>Multi-task w/ Synthetic Supervision</i>										
Target Inference	27.70	28.53	0.83	10.97	7.25	46.64	43.23	3.41	12.40	1.63
Style Inference	26.67	28.65	1.98	<b>14.53</b>	<b>12.58</b>	45.46	44.16	1.30	<b>14.52</b>	<b>2.19</b>

Table 7.3: All FSMT systems achieve better BLEU scores when the intended formality matches the nature of the text being translated (scores are grayed otherwise). LEPoD scores (all scores are percentages) show that synthetic supervision introduces more changes between formal and informal outputs than baselines, and Online Style Inference produces the most diverse informal/formal translations.

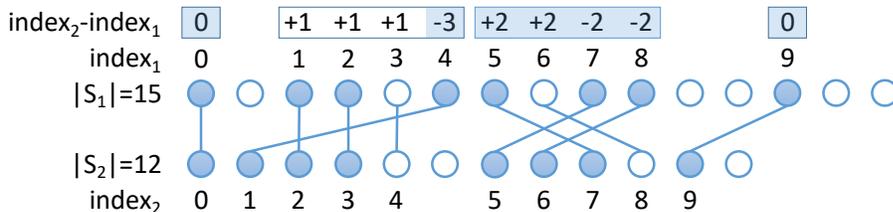


Figure 7.2: Comparing  $S_1$  and  $S_2$  with LEPoD: hollow circles represent non-exact matched tokens, yielding a LED score of  $(\frac{7}{15} + \frac{4}{12}) \times \frac{1}{2} = 0.4$ . Given the alignment illustrated above, the PoD score is  $\frac{0+3+2+0}{10} = 0.5$ .

We first compute the pairwise Lexical Difference (LED) based on the percentages of tokens that are not found in both outputs. Formally,

$$\text{LED} = \frac{1}{2} \left( \frac{|S_1 \setminus S_2|}{|S_1|} + \frac{|S_2 \setminus S_1|}{|S_2|} \right), \quad (7.6)$$

where  $S_1$  and  $S_2$  is a pair of sequences and  $S_1 \setminus S_2$  indicates tokens appearing in  $S_1$  but not in  $S_2$ .

We then compute the pairwise Positional Difference (PoD) by identifying aligned partitions within the compared segments, and computing the maximum distortion within each partition. Word alignments are obtained using the latest

METEOR software (Denkowski and Lavie, 2014), which supports stem, synonym and paraphrase matches in addition to exact matches. In order to find noncrossing partitions that represent linear ordered paraphrases, we first re-index  $N$  aligned units (words or phrase) and calculate distortions as the position differences (i.e.,  $\text{index}_2 - \text{index}_1$  in Figure 7.2). Then we keep a running total of the distortion array  $(d_1, d_2, \dots)$ , and cut off a partition  $p = (d_i, \dots, d_j) \in P$  whenever the accumulation is zero (i.e.,  $\sum p = 0$ ). Now we can define

$$\text{PoD} = \frac{1}{N} \sum_{p \in P} \max(\text{abs}(p)). \quad (7.7)$$

In extreme cases, when the first unit in  $S_1$  is reordered to the last position in  $S_2$ , PoD score approaches 1. When units are aligned without any reordering, each alignment constitutes a partition and PoD equals 0.

**Findings** LEPOD scores measuring the discrepancy between informal and formal outputs of each model in Table 7.3 show that Multi-Task DS-Tag and Multi-Task get similar lexical and positional variability. The benefit of adding formality tags via offline data selection is unclear, which is also suggested in Chapter 6. Online Target Inference gets slightly larger discrepancy on MSLT, while Online Style Inference performs notably differently. Particularly, the latter has much larger positional discrepancy scores, which indicates that it produces more structural diverse sentences. However, larger surface changes are more likely to alter the meaning, and the changes are not guaranteed to be formality-oriented. We therefore use this study to select the most promising models for human evaluation: BLEU and LEPOD scores indicate that Online Style Inference produces the most diverse formal and informal

outputs while roughly preserving BLEU. We select this model for further human evaluation and compare it against Multi-Task.

## 7.5 Human Evaluation of FSMT

**Methodology** We want to directly measure the improvement of Online Style Inference over Multi-Task, so a different human evaluation protocol from Chapter 6 is used here. Our evaluation scheme asks annotators to directly compare sentence pairs on two criteria, *meaning preservation* and *formality difference*, and obtains win:tie:loss ratios.<sup>10</sup>

**Meaning Preservation** We ask annotators to compare outputs of two systems against the reference translation, and decide which one better preserves the meaning of the reference. The following instruction is provided to annotators.

For each task, you will be presented with an English sentence and two rewrites of that sentence. Your task is to judge which rewrite better preserves the meaning of the original and choose from:

- Rewrite 1 is much better
- Rewrite 1 is better
- No preference between Rewrite 1 and Rewrite 2  
(no difference in meaning or hard to say)
- Rewrite 2 is better
- Rewrite 2 is much better

---

<sup>10</sup>We do not evaluate fluency in this chapter because both [Rao and Tetreault \(2018\)](#) and in Chapter 6 we show various automatic systems achieve an almost identical fluency level. Annotators also have systematically biased feeling in fluency when comparing formal and informal sentences as suggested in Chapter 4 and by [Rao and Tetreault \(2018\)](#).

Note that this task focuses on differences in content, so differences in style (such as formality) between the original and rewrites are considered okay. [Some examples with explanations are provided.]

**Formality Difference** We ask annotators to compare outputs of two systems and decide which is more formal. The following instruction is displayed.

People use different varieties of language depending on the situation: formal language is required in news articles, official speeches or academic assignments, while informal language is more appropriate in instant messages or spoken conversations between friends.

You will be presented with two English sentences, and your task is to decide which one is more formal and choose from:

- Sentence 1 is much more formal
- Sentence 1 is more formal
- No preference between Sentence 1 and Sentence 2  
(no difference in formality or hard to say)
- Sentence 2 is more formal
- Sentence 2 is much more formal

Keep in mind:

- Language formality can be affected by many factors, such as the choices of grammar, vocabulary, and punctuation.
- The sentences in the pair could have different meanings. Please rate the formality of the sentences independent of their meaning.
- The sentences in the pair could be nonsensical. Please rate the formality of the sentences independent of their quality.

Generally, a sentence with small formality changes such as fewer contractions, proper punctuation or some formal terms is considered “more formal”. A sentence is considered “much more formal” if it contains multiple indicators of formality, or if the sentence construction itself reflects a more formal style. That said, feel free to use your own judgment for doing the task if what you see is not covered by these examples. [Some examples with explanations are provided.]

We randomly sample  $\sim 150$  examples from WMT and MSLT respectively, and obtain judgments for informal and formal translations of each example. We collect these judgments from 30 volunteers who are native or near-native English speakers. Annotators only compare translations of the same (intended) formality generated by different systems. Identical translation pairs are excluded. Each comparison receives five independent judgments, unless the first three judgments are identical.

The inter-rater agreement using Krippendorff’s alpha is  $\sim 0.5$ . It indicates that there is some variation in annotators’ assessment of language formality. We therefore follow the majority and take the competence of annotators into consideration. In Figure 7.3, independent judgments are aggregated using MACE (Hovy et al., 2013), which estimates the competence for annotators.

**Findings** Overall, the human evaluation shows that synthetic supervision successfully improves desired formality of the output while preserving translation quality, compared to a strong multi-task baseline. Figure 7.3a and 7.3b show that informal translations generated by Online Style Inference are annotated as more informal

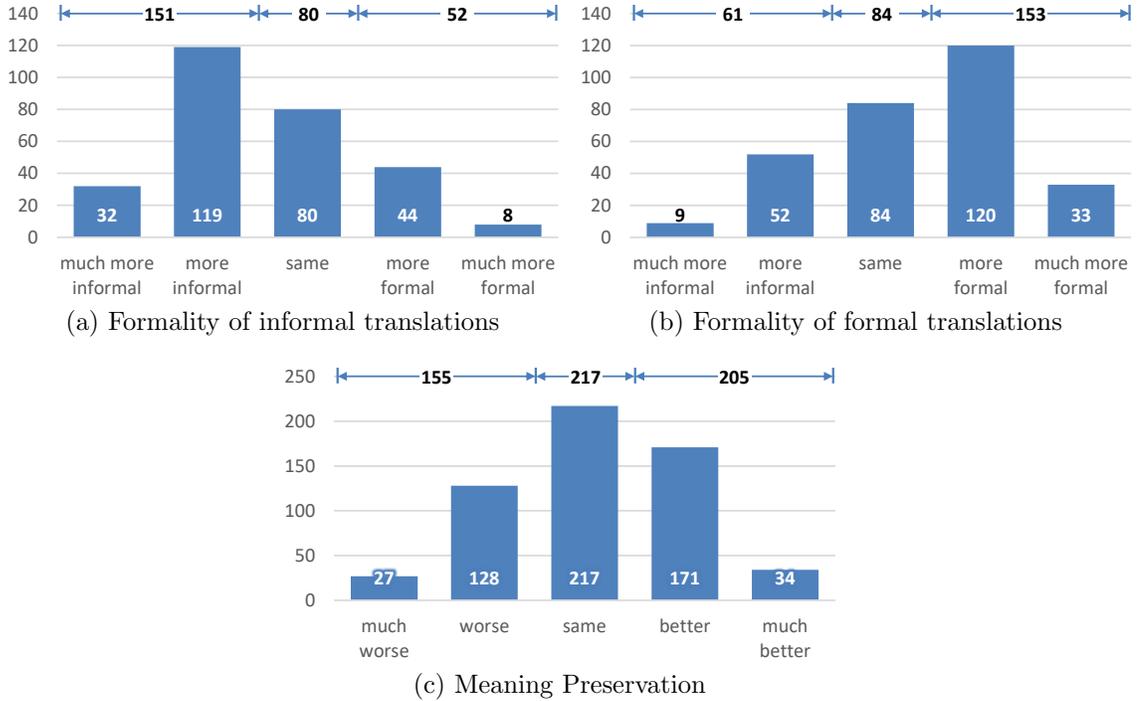


Figure 7.3: Win/Tie/Loss counts when comparing Online Style Inference to Multi-Task. Informal translations generated by OSI are annotated as more informal than Multi-Task, while formal translations are annotated as more formal. The OSI model also gets more instances that better preserve the meaning.

than the baseline model (**win:tie:loss=151:80:52**), while formal translations are annotated as more formal (**win:tie:loss=153:84:61**). For both cases, the win-loss differences are significant with  $p < 0.001$  using the sign test, where ties are evenly distributed to wins and losses as suggested by Demsar (2006). The results confirm that synthetic supervision lets the model better tailor its outputs to the desired formality, and suggest that the differences between formal and informal outputs detected by the LEPoD scores are indeed representative of formality changes. Online Style Inference preserves the meaning of the source better than Multi-Task (**win:tie:loss=205:217:155**), as shown in Figure 7.3c. The win-loss difference for meaning preservation is still significant with  $p < 0.02$ , but less stronger than for-

mality difference.

## 7.6 Qualitative Analysis

We conduct further analysis semi-automatically to better understand how formal and informal translations differ from each other. Most types of changes made by human rewriters (Pavlick and Tetreault, 2016; Rao and Tetreault, 2018) are also observed in our system outputs (examples can be found in Table 7.4).

We first check how often the systems output the same translation for formal and informal style. As can be seen in Table 7.5, both synthetic supervision methods improve over the baseline multi-task system, and the best Online Style Inference system introduces changes between formal and informal translations 12% more often in 6,546 test examples compared to the baseline.

Manual inspection reveals simple patterns indicative of formality changes. We implement rules to check how often these patterns are found in FSMT output (Table 7.5). A sentence can be made more formal by expanding contractions (contr.) and removing unnecessary fillers such as conjunctions (*so/and/but*) and interjections (*well*) at the beginning of a sentence (filler). Online Target Inference performs these changes more frequently. We also examine the introduction of quotation marks in formal translations (quot.); using possessive *of* instead of possessive *'s* (poss.); and rewrites of informal use of declarative form for yes-no questions (y/n). Online Style Inference output matches these patterns better than other systems.

Finally, we conduct a manual analysis to understand the nature of differences

Type	Informal translation	Formal translation
Filler	And I think his wife has family there.	I think his wife has family there.
Completeness ▼		
Quotation	The gas tax is simply not sustainable, said Lee.	“The gas tax is simply not sustainable,” said Lee.
Yes-No	You like shopping?	Do you like shopping?
Subject	Sorry it’s my fault.	I’m sorry it’s my fault.
Article	Cookies where I work.	The cookies where I work.
Relativizer	Other stores you can’t buy.	The other stores where you can’t buy.
Paraphrasing ▼		
Contraction	I think he’d like that, but we’ll see.	I think he would like that, but we will see.
Possessive	Fay’s innovation perpetuated over the years.	The innovation of Fay has perpetuated over the years.
Adverb	I told you already.	I already told you.
Idiom	Hi, how’s it going?	Hi, how are you?
Slang	You gotta let him digest.	You have to let him digest.
Word-1	Actually my dad’s some kind of technician so he understands, but my mom’s very old.	In fact, my father is some kind of technician so he understands, but my mother is very old.
Word-2	Maybe a little more in some areas.	Perhaps a little more in certain areas.
Word-3	It’s really necessary for our nation.	This is essential for our nation.
Phrase-1	Yeah, me neither.	Yeah, neither do I.
Phrase-2	I think he’s moving to California now.	I think he is moving to California at the moment.
Phrase-3	It could be a Midwest thing.	This could be one thing from the Midwest.

Table 7.4: Types of the differences between informal and formal translations. Examples are drawn from the output of Online Style Inference.

Model	identical	contr.	filler	quot.	poss.	y/n	$\Delta$ length
Multi-Task	2,140 (33%)	915	530	146	46	13	1.30
Online Target Inference	1,868 (29%)	<b>1,370</b>	<b>635</b>	145	41	21	1.58
Online Style Inference	<b>1,385 (21%)</b>	1,347	530	<b>252</b>	<b>86</b>	<b>33</b>	<b>4.57</b>

Table 7.5: Heuristic analysis of the differences between informal and formal translations. Both synthetic supervision methods introduce more changes between formal and informal translations. Online Target Inference usually performs simple substitutions while Online Style Inference performs more less-deterministic changes. Online Style Inference also generates more complete and longer formal translations.

between formal and informal translations of Online Style Inference that are not represented by the simple patterns. We observe that ellipsis is frequent in informal outputs, while formal sentences are more complete, using complement subjects, proper articles, conjunctions, relative pronouns, etc. This is reflected in their longer length ( $\Delta$ length in Table 7.5 is the average length difference in characters). Lexical or phrasal paraphrases are frequently used to convey formality, substituting familiar terms with more formal variants (e.g., “grandma” vs. “grandmother”). Examining translations with large POD scores shows that Online Style Inference is more likely to reorder adverbs based on formality: e.g., “I told you already” (I) vs. “I already told you” (F).

A few types of human rewrites categorized by [Pavlick and Tetreault \(2016\)](#) and [Rao and Tetreault \(2018\)](#) are not observed here. For example, our models almost always produce words with correct casing and standard spelling for both informal and formal languages. This matches the characteristics of the translation data we used for training.

We manually inspect system outputs that fail to preserve the source meaning and reveal some limitations of using synthetic supervision. (1) Inaccurate synthetic

labels introduce noise. Online Target Inference sometimes generates “I am not sure” as the formal translation, regardless of the source. We hypothesize that this is due to the imperfect synthetic translations generated by the formality transfer sub-model reinforce this error pattern. (2) Synthetic data may not reflect the true distribution. Occasionally, Online Style Inference drops the first word in a formal sentence even if it is not a filler, e.g. “~~O~~n Thursday, ...” We hypothesize that labeling too many formal/informal examples of similar patterns could lead to ignoring context. While Online Style Inference improves meaning preservation comparatively, it still bears the challenge of altering meaning when fitting to a certain formality, such as generating “there will be no longer than the hill of Runyonyi” when the reference is “then only Rumyoni hill will be left”.

## 7.7 Summary

We explored acquiring synthetic supervision for formality-sensitive machine translation. We introduced two novel approaches that automatically generate synthetic training triples by either inferring the translation from the source sentence and desired formality or inferring the target formality from a given translation pair. Comprehensive automatic and human assessments demonstrated the effectiveness of using synthetic supervision. Our best model outperformed strong baselines by producing translations that better match desired formality levels while preserving the source meaning.

## Chapter 8: Conclusion and Future Work

### 8.1 Summary

This dissertation addressed the problem of automatic formality transfer within and across languages. By modeling style transfer within a language and machine translation jointly, we designed models that are able to generate language for a desired formality level despite limited training data.

We first confirmed a research hypothesis that formality variations for language generation could be modeled from examples, such as a pool of formal and informal words or sentence pairs. We presented an approach to inducing a stylistic subspace using lexical paraphrases and building a formality scorer using representative words. This approach better distinguished more formal from less formal words than using the original space and enabled us to place sentences on a continuous formality scale based on lexical scores (Niu and Carpuat, 2017). We brought the formality model to real-world scenarios and introduced a new task — Formality-Sensitive Machine Translation (FSMT). Given the formality score of sentences, we realized the first formality-constrained language generation system to perform French-English FSMT. It was built based on a standard PBMT architecture and trained only with translation pairs. For each input sentence, the resulting system produces translation

hypotheses of different formality levels and promotes hypotheses whose formality scores are closer to the desired formality level. This system can effectively control language formality (Niu et al., 2017).

The space of possible outputs of the PBMT-based system is limited to lexical changes and  $n$ -best translation hypotheses, so we turned to using neural models to capture more context. We started by using neural sequence-to-sequence models for directly modeling the formality variation at the sentence level, i.e., formality transfer. Since informal-formal sentence pairs are only available in limited quantity, we took a detour and researched a related problem — improving NMT quality in low-resource settings. We designed a bi-directional NMT framework that jointly translates in both translation directions with a single model. It can be used for efficient iterative back-translation since no auxiliary models are required (Niu et al., 2018a). We also introduced a differentiable input reconstruction loss for it to exploit the source side of parallel samples without additional parameters (Niu et al., 2019). The bi-directional NMT framework was then successfully applied to monolingual formality transfer tasks (Niu et al., 2018b).

Afterward, we handled tasks of formality style transfer within and across languages altogether and confirmed that models of these two tasks could help each other. We built a neural system by jointly training on both formality transfer and machine translation data. The joint training yielded a single model that not only significantly improves the quality of formality transfer for English in both directions but also performs FSMT without being explicitly trained on style-annotated translation examples (Niu et al., 2018b). This neural FSMT system provides higher

formality-control intensity than the PBMT-based system, but sometimes produces translations disobeying the source meaning. Finally, we introduced training with synthetic supervision to further improved the performance of the neural FSMT system. A formality transfer submodule embedded in the joint model was used to infer the target formality from a given translation pair. After being trained with complete training triplets, the FSMT system produces translations that better match desired formality levels while preserving the source meaning (Niu and Carpuat, 2019).

## 8.2 Future Work

To wrap up this dissertation, we discuss limitations and directions for future research.

### 8.2.1 Modeling Formality in the Neural Architecture

While effective, neural formality transfer models are opaque. When working with PBMT, we explicitly modeled lexical formality. That enabled us to identify which words make a sentence more informal or more formal. However, after tuning to the neural architecture, we relied on the heavily parameterized neural network to contextually make appropriate word choices when generating a sentence at the desired formality level. We would like to reveal the underlying mechanism of how these choices were made. For example, we could discover what contextual words were indicative of style by analyzing model decisions using the attention (e.g., Xu et al., 2015; Ghaeini et al., 2018) or gradients (e.g., Feng et al., 2018; Jain and

Wallace, 2019).

Our neural FSMT model only takes a binary view of formal-informal distinctions, which limits the granularity of formality control. By contrast, the PBMT-based system is capable of promoting translations of different formality levels as it can take numerical formality scores as input. In future work, we would like to relax the constraint of using only opposite styles. We could start from using interpolation for formality tags’ representations, i.e., creating a new target formality embedding by a linear combination  $w \cdot \langle \text{Formal} \rangle + (1 - w) \cdot \langle \text{Informal} \rangle$  (Johnson et al., 2017).

## 8.2.2 A Broader Range of Tasks

We focused on controlling the target formality in two language generation tasks — monolingual formality transfer and FSMT. There are other interesting tasks worth exploring.

(1) In the FSMT task, instead of providing desired formality, we would also like to infer the source formality to preserve the formality level. This scenario is helpful if the user is interested in the source style. However, a challenge of achieving this lies in the difficulty of aligning formality levels between two languages: one sentence may not have an equivalent preserving the formality in the other language. For example, the French sentence “vous êtes arrivé” could be translated into English “you have arrived”, but the information that “vous” is a formal second person singular pronoun has lost. Alternatively, from English to French, it is unclear whether “vous” or “tu”

(i.e., an informal second person singular pronoun) preserves the formality level of “you”.

(2) Although formality is considered a key dimension of style, modeling other styles (e.g., complexity and specificity) is also desired in practice (Chandrasekar et al., 1996; Enç, 1991). We would like to investigate whether our proposed methods are also effective on other styles and address potential issues not revealed when dealing with the formality. For example, text written for a higher reading grade level often includes more details than a simpler version aimed at a lower reading grade level. As a result, making a simple sentence more complex might require adding content that was not present in the input sentence.

(3) We would like to broaden our horizon to a variety of language generation tasks that will benefit from constraining certain styles, such as dialog generation (Li et al., 2016) and poetry generation (Zhang and Lapata, 2014). In these scenarios, the input is not iterated in other words or languages, so how to make use of the style transfer data (i.e., paraphrasing data) remains challenging.

### 8.2.3 Challenges of Joint Training

We would like to further relax the assumptions on the nature of data available to build FSMT systems. We built the neural FSMT system by training on a concatenation of the formality transfer and machine translation data. There is a domain mismatch between these two datasets since the formality transfer data comes from Yahoo Answers while the machine translation data mostly comes from Open-

Subtitles. We attribute a portion of inaccurate formality-controlled translations to this issue. Selecting a subset of formality transfer examples that are closer to the MT data is not advised because they are intrinsically insufficient. We would like to borrow or explore techniques for unsupervised style transfer or unsupervised NMT (Artetxe et al., 2018; Lample et al., 2018a,b; Wu et al., 2019) to adapt formality transfer models to the domain of MT by leveraging unpaired data. This research direction would make it possible to model other types of style variations and target languages where transfer examples are not readily available.

We only considered French to English for the FSMT task, and the joint model uses a shared encoder for both French and English input. French and English are from the same language family (i.e., Indo-European), which makes vocabulary and embedding space sharing relatively easy. We would like to experiment with more language pairs that are more distant and draw from advances in multilingual NMT literature to handle potential issues.

#### 8.2.4 Differentiable Sampling for Unsupervised and Semi-Supervised Training

We briefly stepped into low-resource NMT and introduced differentiable reconstruction from sampled sequences. This technique has the potential to be useful for many unsupervised and semi-supervised sequence generation tasks. For example, we could apply the round-trip translation to monolingual corpora in addition to parallel corpora for NMT. We would also like to bring this technique to style-constrained lan-

guage generation tasks. For example, we could target a specific property that needs to be improved (e.g., style intensity and output fluency) and design a differentiable loss that evaluates sampled sequences.

## Bibliography

- Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1197. Association for Computational Linguistics.
- Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roei Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *CoRR*, abs/1903.07091.
- Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.
- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations*.
- Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the Third International Conference on Learning Representations*.
- Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don’t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247. Association for Computational Linguistics.
- Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, abs/1308.3432.
- Adam L. Berger, Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, John R. Gillett, John D. Lafferty, Robert L. Mercer, Harry Printz, and Lubos

- Ures. 1994. The candid system for machine translation. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*. Morgan Kaufmann.
- Douglas Biber. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.
- Douglas Biber. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, 14(1):7–34.
- Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press.
- Graeme W. Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3112–3122. Association for Computational Linguistics.
- Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214. Association for Computational Linguistics.
- Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation*, pages 272–303. Association for Computational Linguistics.
- Richard W. Brislin. 1970. Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1(3):185–216.
- Julian Brooke and Graeme Hirst. 2014. Supervised ranking of co-occurrence profiles for acquisition of continuous lexical attributes. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 2172–2183. Dublin City University and Association for Computational Linguistics.
- Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics, Posters Volume*, pages 90–98. Chinese Information Processing Society of China.
- Penelope Brown and Colin Fraser. 1979. Speech as a marker of situation. In *Social Markers in Speech*, chapter 2, pages 33–62. Cambridge University Press.
- Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

- Roger Brown and Albert Gilman. 1960. The pronouns of power and solidarity. In *Style in Language*, pages 253–276. MIT Press.
- Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the common crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3579–3584. European Language Resources Association.
- Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media*.
- Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the Bible. *Royal Society Open Science*, 5(10).
- Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.
- Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsutho Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 2–14.
- Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 1041–1044.
- Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1965–1974. Association for Computational Linguistics.
- Colin Cherry and George F. Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436. Association for Computational Linguistics.
- Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, 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*, pages 1724–1734. Association for Computational Linguistics.
- Jihun Choi, Kang Min Yoo, and Sang-goo Lee. 2018. Learning to compose task-specific tree structures. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5094–5101. AAAI Press.
- Christos Christodoulopoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

- Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. European Language Resources Association.
- Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181. Association for Computational Linguistics.
- Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167. ACM.
- Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.
- William Coster and David Kauchak. 2011. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9. Association for Computational Linguistics.
- Ryan Cotterell and Julia Kreutzer. 2018. Explaining and generalizing back-translation through wake-sleep. *CoRR*, abs/1806.04402.
- Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991.
- Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.
- Janez Demsar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.
- Misha Denil, Babak Shakibi, Laurent Dinh, Marc’Aurelio Ranzato, and Nando de Freitas. 2013. Predicting parameters in deep learning. In *Advances in Neural Information Processing Systems 26*, pages 2148–2156. Curran Associates, Inc.
- Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380. Association for Computational Linguistics.
- Chrysanne DiMarco and Keith Mah. 1994. A model of comparative stylistics for machine translation. *Machine Translation*, 9(1):21–59.

- Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1723–1732. Association for Computational Linguistics.
- Susan T Dumais, George W Furnas, Thomas K Landauer, Scott Deerwester, and Richard Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 281–285. ACM.
- Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648. Association for Computational Linguistics.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500. Association for Computational Linguistics.
- Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *Proceedings of the International Conference on Language Resources and Evaluation*. European Language Resources Association.
- Mürvet Enç. 1991. The semantics of specificity. *Linguistic Inquiry*, 22(1):1–25.
- Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan L. Boyd-Graber. 2018. Pathologies of neural models make interpretation difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728. Association for Computational Linguistics.
- 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. Association for Computational Linguistics.
- Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875. Association for Computational Linguistics.
- Eric N. Forsyth and Craig H. Martell. 2007. Lexical and discourse analysis of online chat dialog. In *Proceedings of the First IEEE International Conference on Semantic Computing*, pages 19–26. IEEE Computer Society.

- Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 663–670. AAAI Press.
- Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29*, pages 1019–1027. Curran Associates, Inc.
- Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 848–856. Association for Computational Linguistics.
- Leon A. Gatys, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. 2016a. Preserving color in neural artistic style transfer. *CoRR*, abs/1606.05897.
- Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016b. Image style transfer using convolutional neural networks. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423. IEEE Computer Society.
- Tao Ge, Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. Automatic grammatical error correction for sequence-to-sequence text generation: An empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6059–6064. Association for Computational Linguistics.
- Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.
- Reza Ghaeini, Xiaoli Z. Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4957. Association for Computational Linguistics.
- John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of IEEE International Conference on Speech, and Signal Processing*, pages 517–520. IEEE.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41. Association for Computational Linguistics.
- Jiatao Gu, Daniel Jiwoong Im, and Victor O. K. Li. 2018. Neural machine translation with gumbel-greedy decoding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5125–5132. AAAI Press.

- Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2017. Effective strategies in zero-shot neural machine translation. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 105–112.
- Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1689–1701. Association for Computational Linguistics.
- Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.
- Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems 30*, pages 2146–2156. Curran Associates, Inc.
- Samuel Ichiye Hayakawa. 1994. *Choose the right word*. Collins Reference.
- Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29*, pages 820–828. Curran Associates, Inc.
- Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.
- Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Internal Report, Center “Leo Apostel”, Free University of Brussels*.
- Francis Heylighen and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7(3):293–340.
- Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.
- Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24. Association for Computational Linguistics.

- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.
- Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130. Association for Computational Linguistics.
- Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.
- 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. PMLR.
- Markus Hurtig. 2006. *Varieties of English in the Swedish classroom*. Karlstad University: Unpublished C-Essay.
- Judith T. Irvine. 1979. Formality and informality in communicative events. *American Anthropologist*, 81(4):773–790.
- 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*, pages 3543–3556. Association for Computational Linguistics.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *Proceedings of the 5th International Conference on Learning Representations*.
- Harsh Jhamtani, Varun Gangal, Eduard H. Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19. Association for Computational Linguistics.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. 2016. The AMU-UEDIN submission to the WMT16 news translation task: Attention-based NMT

- models as feature functions in phrase-based SMT. In *Proceedings of the First Conference on Machine Translation*, pages 319–325. Association for Computational Linguistics.
- Tomoyuki Kajiwara. 2019. Negative lexically constrained decoding for paraphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6047–6052. Association for Computational Linguistics.
- Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709. Association for Computational Linguistics.
- Jack Kiefer and Jacob Wolfowitz. 1952. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466.
- 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. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3th International Conference on Learning Representations*.
- Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, System Demonstrations*, pages 67–72. Association for Computational Linguistics.
- Philipp Koehn. 2004a. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Machine Translation: From Real Users to Research, 6th Conference of the Association for Machine Translation in the Americas*, volume 3265 of *Lecture Notes in Computer Science*, pages 115–124. Springer.
- Philipp Koehn. 2004b. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395. Association for Computational Linguistics.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 79–86.
- Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

- Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the 2005 International Workshop on Spoken Language Translation*, pages 68–75. International Speech Communication Association.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics.
- Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39. Association for Computational Linguistics.
- Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133. Association for Computational Linguistics.
- Elizaveta Korotkova, Maksym Del, and Mark Fishel. 2018. Monolingual and cross-lingual zero-shot style transfer. *CoRR*, abs/1808.00179.
- Shibamouli Lahiri. 2015. SQUINKY! A corpus of sentence-level formality, informativeness, and implicature. *CoRR*, abs/1506.02306.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations*.
- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049. Association for Computational Linguistics.
- Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc’Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *Proceedings of the Seventh International Conference on Learning Representations*.
- Ekaterina Lapshinova-Koltunski and Mihaela Vela. 2015. Measuring ‘registerness’ in human and machine translation: A text classification approach. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 122–131. Association for Computational Linguistics.

- Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180. Association for Computational Linguistics.
- Will Lewis, Christian Federmann, and Ying Xin. 2015. Applying cross-entropy difference for selecting parallel training data from publicly available sources for conversational machine translation. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, pages 126–134.
- Haiying Li, Zhiqiang Cai, and Arthur C. Graesser. 2013. Comparing two measures for formality. In *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference*, pages 220–225. AAAI Press.
- Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 994–1003. Association for Computational Linguistics.
- Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. European Language Resources Association.
- Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921. Association for Computational Linguistics.
- Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys*, 40(3):8:1–8:49.
- Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *Proceedings of the 4th International Conference on Learning Representations*.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics.
- Chris J. Maddison, Daniel Tarlow, and Tom Minka. 2014. A\* sampling. In *Advances in Neural Information Processing Systems 27*, pages 3086–3094. Curran Associates, Inc.

- François Mairesse and Marilyn A. Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488.
- Daniel Marcu and Daniel Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 133–139. Association for Computational Linguistics.
- David D. McDonald and James Pustejovsky. 1985. A computational theory of prose style for natural language generation. In *Proceedings of the Second Conference of the European Chapter of the Association for Computational Linguistics*, pages 187–193. Association for Computational Linguistics.
- 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*, pages 312–318. Association for Computational Linguistics.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the First International Conference on Learning Representations, Workshop Track*.
- Hideki Mima, Osamu Furuse, and Hitoshi Iida. 1997. Improving performance of transfer-driven machine translation with extra-linguistic information from context, situation and environment. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 983–989. Morgan Kaufmann.
- Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 495–504. Association for Computational Linguistics.
- Robert C. Moore and William D. Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 220–224. Association for Computational Linguistics.
- Jonas Mueller, David K. Gifford, and Tommi S. Jaakkola. 2017. Sequence to better sequence: Continuous revision of combinatorial structures. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2536–2544. PMLR.
- Toan Q. Nguyen and David Chiang. 2018. Improving lexical choice in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 334–343. Association for Computational Linguistics.

- Eugene Albert Nida and Charles Russell Taber. 2003. *The Theory and Practice of Translation*, volume 8 of *Helps for Translators*. Brill.
- Xing Niu and Marine Carpuat. 2017. Discovering stylistic variations in distributional vector space models via lexical paraphrases. In *Proceedings of the Workshop on Stylistic Variation*, pages 20–27. Association for Computational Linguistics.
- Xing Niu and Marine Carpuat. 2019. Controlling neural machine translation formality with synthetic supervision. *Under Review*.
- Xing Niu, Michael Denkowski, and Marine Carpuat. 2018a. Bi-directional neural machine translation with synthetic parallel data. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 84–91. Association for Computational Linguistics.
- Xing Niu, Marianna J. 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. Association for Computational Linguistics.
- Xing Niu, Sudha Rao, and Marine Carpuat. 2018b. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021. Association for Computational Linguistics.
- Xing Niu, Weijia Xu, and Marine Carpuat. 2019. Bi-directional differentiable input reconstruction for low-resource neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 442–448. Association for Computational Linguistics.
- Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167. Association for Computational Linguistics.
- Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alexander M. Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir R. Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the 2004 Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting*, pages 161–168. Association for Computational Linguistics.
- Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302. Association for Computational Linguistics.

- Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.
- Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.
- Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Association for Computational Linguistics.
- Myle Ott, Michael Auli, David Grangier, and Marc’Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3953–3962. PMLR.
- Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 560–569. Association for Computational Linguistics.
- 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 of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.
- Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224. Association for Computational Linguistics.
- Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 425–430. Association for Computational Linguistics.
- Ellie Pavlick and Joel R. Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.
- Karl Pearson. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572.
- Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom M. Mitchell. 2018. Contextual parameter generation for universal neural machine

- translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435. Association for Computational Linguistics.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation*, pages 186–191. Association for Computational Linguistics.
- Richard Power, Donia Scott, and Nadjat Bouayad-Agha. 2003. Generating texts with style. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, volume 2588 of *Lecture Notes in Computer Science*, pages 444–452. Springer.
- 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*, pages 866–876. Association for Computational Linguistics.
- Daniel Preotiuc-Pietro, Wei Xu, and Lyle H. Ungar. 2016. Discovering user attribute stylistic differences via paraphrasing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3030–3037. AAAI Press.
- Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 157–163. Association for Computational Linguistics.
- Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, 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*, pages 1074–1084. Association for Computational Linguistics.
- 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*, pages 129–140. Association for Computational Linguistics.
- Ehud Reiter and Sandra Williams. 2010. Generating texts in different styles. In *The Structure of Style - Algorithmic Approaches to Understanding Manner and Meaning*, pages 59–75. Springer.
- Jack C. Richards, John Platt, and Heidi Platt. 1997. *Longman Dictionary of Language Teaching and Applied Linguistics*. Longman.
- Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

- Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777. Association for Computational Linguistics.
- Sascha Rothe and Hinrich Schütze. 2016. Word embedding calculus in meaningful ultradense subspaces. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 512–517. Association for Computational Linguistics.
- Natalie Schilling-Estes. 2002. Investigating stylistic variation. In *The Handbook of Language Variation and Change*, chapter 15, pages 375–401. Blackwell Publishing Ltd.
- Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 83–91. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics.
- Fadi Abu Sheikha and Diana Inkpen. 2011. Generation of formal and informal sentences. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 187–193. Association for Computational Linguistics.
- Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the 2004 Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting*, pages 177–184. Association for Computational Linguistics.
- Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30*, pages 6833–6844. Curran Associates, Inc.

- Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1374–1383. Association for Computational Linguistics.
- Zhiyi Song, Stephanie M. Strassel, Haejoong Lee, Kevin Walker, Jonathan Wright, Jennifer Garland, Dana Fore, Brian Gainer, Preston Cabe, Thomas Thomas, Brendan Callahan, and Ann Sawyer. 2014. Collecting natural SMS and chat conversations in multiple languages: The BOLT phase 2 corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1699–1704. European Language Resources Association.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.
- Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3097–3103. AAAI Press.
- Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103. ACM.
- Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1503–1515. Association for Computational Linguistics.
- Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018a. Translating pro-drop languages with reconstruction models. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 4937–4945. AAAI Press.
- Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2018b. Learning to jointly translate and predict dropped pronouns with a shared reconstruction mechanism.

- In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2997–3002. Association for Computational Linguistics.
- Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. 2016. Text simplification using neural machine translation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 4270–4271. AAAI Press.
- Yijun Wang, Yingce Xia, Li Zhao, Jiang Bian, Tao Qin, Guiquan Liu, and Tie-Yan Liu. 2018c. Dual transfer learning for neural machine translation with marginal distribution regularization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5553–5560. AAAI Press.
- Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2016. Measuring the effect of conversational aspects on machine translation quality. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 2571–2581. The COLING 2016 Organizing Committee.
- Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J. 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. Association for Computational Linguistics.
- Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280.
- Jiawei Wu, Xin Wang, and William Yang Wang. 2019. Extract and edit: An alternative to back-translation for unsupervised neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1173–1183. Association for Computational Linguistics.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.
- Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 1015–1024. Association for Computational Linguistics.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and

- tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057. PMLR.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.
- Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2899–2914. The COLING 2012 Organizing Committee.
- 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*, pages 203–210. The COLING 2016 Organizing Committee.
- Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.
- Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.
- Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594. Association for Computational Linguistics.
- Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 555–562. AAAI Press.
- Zheming Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1353–1361. The COLING 2010 Organizing Committee.