Parallel Training Data Selection for Conversational Machine Translation

Xing Niu, Marine Carpuat

Dept. of Computer Science
University of Maryland, College Park
{xingniu,marine}@cs.umd.edu

Abstract

We describe data selection strategies for an English-French Skype conversation translation task without in-domain training data. Selection methods based on language modeling and text formality criterion are evaluated. Our main finding is that translating conversation transcripts turned out to not be as challenging as we expected: while translation quality is of course not perfect, a straightforward phrase-based system trained on movie subtitles yields high BLEU scores, and small improvements are obtained by using a simple heuristic to select more Skype-like examples.

1. Introduction

Our goal was to evaluate the strength and weaknesses of various parallel data selection strategies for translating conversational transcripts. We focus on an English-French Skype conversation translation task (MSLT at 2016 IWSLT evaluation campaign [1]). This is a challenging task for several reasons. First, only small amounts of conversation transcripts (45k words of English) are available for system development. Second, we anticipate that Skype conversations diverge from the vast majority of available training corpora along many dimensions, including differences in what is talked about (content, topics) and how it is talked about (style, register). Our primary task in building a Machine Translation (MT) system will therefore be to identify useful training examples from a large pool of diverse out-of-domain parallel corpora.

Our MT systems uses a standard phrase-based architecture, outlined in Section 2. Our experiments focused on determining what kind of training data is most useful to translate this compared various of data selection techniques (Sections 4) to make the most of the available data (Section 3), and finally conduct an error analysis (Section 5).

Translating such conversation transcripts in these settings turned out to not be as challenging as we expected: while translation quality is of course not perfect, a straightforward phrase-based system trained on movie subtitles yields high BLEU scores (high 40s on the development set) and manual analysis of 100 examples showed that 61 of them were correctly translated, and errors were mostly local disfluencies in the remaining examples. Small improvements in BLEU were obtained by using a simple heuristic to select more Skype-like examples. This approach was the only data selection approach that improved performance. Using language modeling and text formality criterion to select examples that we expected to be closer to Skype data did not improve BLEU.

2. Machine Translation Architecture

2.1. Core Configuration

We use the `Moses` [2] statistical machine translation toolkit to build phrase-based MT systems. Training the MT systems was mostly done by following the standard `Moses` pipeline[1] with default parameters: the maximum length of phrases was limited to 7 words, reordering was limited to 6 words skipped and the reordering model was specified as *msd-bidirectional-fe*. Word alignments were generated using `fast_align` [3], and symmetrized using the *grow-diag-final-and* heuristic. We used 4-gram language models, trained using `KenLM` [4] with modified Kneser-Ney smoothing [5]. Model weights were tuned using the MERT algorithm [6] with 5-fold cross validation.

2.2. Model Weights and Decoding

In order to make the most of the data available during development time, we split the MSLT-dev corpus into five parts of equal size. Log-linear model weights were tuned on all four-combinations of these subsets.

During the system development phase, we use this split to evaluate our systems with 5-fold cross-validation. Decoders with distinct sets of weights are used to decode each of the dev-set splits, and their outputs were concatenated as a whole to be evaluated by the BLEU score. All experiments mentioned in this paper were evaluated in this way.

For the evaluation, we used the five sets of model weights differently: we averaged the weights to create a new set of model weights that was used to decode the evaluation test set. We did verify that weight averaging improved BLEU on the development data.

3. Data Preparation

The MSLT task provided no in-domain training data, but a number of English-French parallel corpora from various genres were permissible. We experimented most of them ex-

[1] `http://www.statmt.org/moses/?n=Moses.Baseline`

cept United Nations parallel corpus because it is analogous to MultiUN and the UN corpora are most dissimilar to Skype calls. The remaining corpora were obtained from translation tasks at the WMT 2016 [2] . Table 1 lists the data statistics. It is worth mentioning that only MSLT and OpenSubtitles data have average sentence lengths of less than 10.

Corpus	# Sentences	# Words (en/fr)
OpenSubtitles	33.5 M	284.0 M / 268.3 M
MultiUN	13.2 M	367.1 M / 432.3 M
Common Crawl	3.2 M	81.1 M / 91.3 M
Europarl v7	2.0 M	55.7 M / 61.9 M
Wikipedia	396 k	9.7 M / 8.7 M
TED corpus	207 k	4.5 M / 4.8 M
News Commentary v10	199 k	5.1 M / 6.3 M
MSLT-dev	5,292	44,865 / 49,562
MSLT-tst	4,854	45,316 / -----

Table 1: English-French parallel data statistics.

As standard pre-processing steps, the parallel data was first tokenized with the Europarl tokenizer[3] and then lowercased with the `Moses` script.

4. Training Data Selection

The goal of the MSLT is to translate the manual transcripts of Skype calls, but the in-domain data is unavailable. Therefore, selecting the most in-domain-like (Skype-like) parallel data was our primary challenge. We tried several strategies to address this problem such as: (1) perplexity-based methods that select most likely sentences/corpora against the lexical probability distribution of the in-domain corpus; (2) formality-based methods that select the sentences at the similar formality level of the in-domain corpus.

4.1. Language Modeling Criterion

We start with well-established data selection techniques for ranking and selecting sentences from the corpora pool.

Using the standard **perplexity**-based selection method [7, 8], the sentences in the out-of-domain corpus are sorted by their perplexity score according to the language model of the in-domain corpus. Suppose the a sentence W consists of N words, its perplexity is defined by:

$$\text{Perplexity}(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}}$$

In-domain-like sentences be retained if they have lowest perplexity score.

Cross-entropy difference is another measurement of corpus similarity [9, 10]. We first trained an in-domain language model LM_{IN} and another language model on the full pool of out-of-domain corpus LM_{OUT}. The algorithm then used these language models to assign a cross-entropy difference score (CED) to each sentence W:

$$H_C(W) = -\frac{1}{N} \log P(w_1 w_2 \dots w_N | C)$$
$$\text{CED}(W) = H_{IN}(W) - H_{OUT}(W)$$

where $H_C(W)$ is the cross-entropy of W on corpus C. Lower scores for cross-entropy difference indicate sentences that are simultaneously more similar to the in-domain corpus and less similar to the full pool average. Lewis et al. [11] also used this approach to select conversational data by using the English Fisher corpus as representative of the domain of interest [12].

4.2. Text Formality Criterion

We anticipate that Skype conversations differ from other training corpora along many content and style dimensions that might be conflated by the language modeling criteria above. We propose to isolate one dimension by selecting data using a text formality criterion. Formality is an important dimension of text stylistic variations [13], and previous work suggests that it can be quantified reliably using minimal supervision [14]. This property is particular interesting in our scenario because Skype conversations is a typical informal register while considerable amount of out-of-domain training data (e.g. MultiUN) stands in the formal side. We therefore introduce formality as a potential data selection metrics via ranking out-of-domain sentences by their informality.

We followed a formality scoring method proposed by [15]. This data-driven model maps a word w to a continuous score via:

$$\text{Formality}(w) = \log \frac{P(w|\,\text{REF})}{P(w|\,\text{ALL})}$$

where REF is the reference corpus and ALL is the combination of all corpora. Specifically, we used MultiUN as the reference of formal language. Word probabilities are estimated by unigram language models with Laplace smoothing. The formality score for a sentence is simply the average score of its words.

Corpus	Formality
MSLT-dev	-2.374
OpenSubtitles	-2.658
TED corpus	-1.621
Common Crawl	-1.254
Wikipedia	-1.199
News Commentary v10	-1.035
Europarl v7	-0.953
MultiUN	-0.508

Table 2: Median sentences' formality scores of different corpora (English part). Higher scores indicate more formal.

Table 2 provides an overview of the median sentence formality scores for different corpora, measured on the English

[2] `http://www.statmt.org/wmt15/translation-task.html`

[3] `http://www.statmt.org/europarl/v7/tools.tgz`

side. A higher score means a corpus being closer to REF (i.e. MultiUN). These distinctions match our intuitions that spoken language (e.g. movie dialogs and TED talks) is more distant from formal than the language used in government proceedings or written news commentary (e.g. Europarl and News).

Formality varies across sentences within the pool of corpora, even within a given corpus, therefore we aimed at selecting sentences with formality scores near a target score. In our scenario, the in-domain corpus (MSLT-dev) got a median score of -2.374, which was the value we targeted. Now we can formally define the third sentence ranking strategy as **formality difference**: the absolute difference between the sentence formality score and the median in-domain formality score.

4.3. Corpus Selection

Since the Skype data for training is unavailable, we selected OpenSubtitles as a pseudo in-domain corpus because it is most similar to Skype conversations by means of both formality and perplexity.

On one hand, Table 2 shows that MSLT-dev and OpenSubtitles has closest formality scores. On the other hand, we also evaluated each sentence by perplexity against a small English language model trained on MSLT-dev. Among all corpora, OpenSubtitles achieves the lowest median sentence perplexity. We only compared the English part because only English sentences are visible when testing.

Training set	BLEU
OpenSubtitles	47.32
+ Wikipedia	**47.70**
+ Wikipedia + TED	47.67
+ TED	47.61
+ Europarl	47.54
+ News_Commentary	47.49
+ Common_Crawl	47.25

Table 3: Translation quality on MSLT-dev when concatenating different corpora with OpenSubtitles as the training data. The BLEU score is computed on uncased tokenized segments.

As listed in Table 3, we also concatenated different corpora with OpenSubtitles as the training data. It suggests that OpenSubtitles+Wikipedia is the best combination and we considered it as our baseline. We skipped concatenating MultiUN or combining all parallel data at hand, because of the tremendous discrepancy between UN documents and Skype conversations.

4.4. Data Selection for Translation Models

As indicated by Table 3, we failed to benefit further from concatenating TED data to Wikipedia. It is not surprising because those corpora are drawn from various domains and genres that are very different from Skype calls. To make use of other out-of-domain corpora, we aimed at selecting a subset that most similar to MSLT-dev using perplexity-based and formality-based methods.

We selected sentence pairs by the English side again because of the same reason that only English sentences are available when testing. Data selection methods rely on the word distributions, but MSLT-dev is far from sufficient to be used for estimating them. Instead, we used OpenSubtitles, the pseudo in-domain corpus, to mimic MSLT-dev.

Method	# Sentences	BLEU
Baseline	34 M	**47.70**
Perplexity	+ 500 k	47.55
Perplexity	+ 1 M	47.65
Cross-entropy diff	+ 500 k	47.43
Cross-entropy diff	+ 1 M	47.48
Formality diff	+ 500 k	47.54
Formality diff	+ 1 M	47.47

Table 4: Translation quality on MSLT-dev when concatenating selected data with OpenSubtitles+Wikipedia as the training data. The BLEU score is computed on uncased tokenized segments.

We ranked sentences from the combination of all available out-of-domain corpora except Wikipedia. 500k or 1M selected parallel sentences were concatenated with our baseline training data. As shown in Table 4, additional OpenSubtitles-like data did not improve the translation quality in this scenario.

4.5. Data Selection for Language Models

A portion of translations using our baseline system encountered the lack of fluency, which suggests a better language model. All experiments so far trained the language models on the same training data for translation models.

Method	# Sentences	BLEU
Baseline	34 M	**47.70**
All data	+ 19 M	47.68
Perplexity	+ 500 k	47.55
Perplexity	+ 1 M	47.66
Cross-entropy diff	+ 500 k	47.61
Cross-entropy diff	+ 1 M	47.61

Table 5: Translation quality on MSLT-dev when concatenating selected data with OpenSubtitles+Wikipedia for training language models. The BLEU score is computed on uncased tokenized segments.

We first used all French sentence from parallel data we have to train a large language model, however, this language model did not improve the BLEU score (see Table 5). Then we utilized the same data selection strategies to select

500k or 1M OpenSubtitles-like French sentences from out-of-domain corpora, and concatenated them to the baseline training data. Unfortunately, none of them helped.

4.6. Heuristic Data Selection

While the formality criterion defined above did not help in selecting useful additional training examples, we notice that on development data, the MT system often incorrectly translates the singular second person pronoun "you" as the formal "vous" in French when the reference translation uses the informl form "tu". We also observe frequent interrogative sentences in the development data. We therefore used these as heuristic rules to select parallel sentences where the French side contains the special tokens: *"tu"* and *"?"*.

Rules	# Sentences	BLEU
Baseline	34 M	47.70
+ *tu*	+ 11 k	47.77
+ *tu, ?*	+ 242 k	**47.82**
+ informal words	+ 385 k	47.44

Table 6: Translation quality on MSLT-dev when concatenating selected data containing special tokens with OpenSubtitles+Wikipedia as the training data. The BLEU score is computed on uncased tokenized segments.

Table 6 shows that these artificially selected data yield a small improvement in translation quality. While the BLEU delta is small, this result suggests that there are indeed useful training examples in the pool of out-of-domain data, but standard data selection techniques failed to capture them successfully.

We attempted to extend this approach by automatically identifying other indicators of informality beyond "tu". Based on the observations that *tu* has an equivalent formal form *vous* and these two words distribute significantly differently in formal/informal corpora, we collected more word pairs that follow these properties.

Specifically, French words w_f and w_i are considered as a formal-informal word pair, if (1) the phrase-table contains two high-scored mappings "$w_e \rightarrow w_f$" and "$w_e \rightarrow w_i$" where w_e is an English word; (2) $\text{Formality}(w_f)$ and $\text{Formality}(w_i)$ are sufficiently different. As a result, 70 informal words were identified, and 385k French sentences containing them were selected.

Unfortunately, the selected data failed to improve BLEU as shown in Table 6. Two possible explanations are: (1) the reference translation does not necessarily always use informal words; (2) the informal words we identified automatically include some reasonable cases such as *ton* (*your* in English) but these are outweighted by noisy examples.

4.7. Sub-sampling for Translation Models

We now turn to a different question. While we have treated the entire OpenSubtitles corpus as a pseudo in-domain corpus so far, are subsets of the corpus more relevant to Skype conversations than others?

To answer this questions, we selected the top half of OpenSubtitles sentences using different ranking strategies (formality difference and perplexity according to the language model trained on MSLT-dev), and compared them with random selection.

Method	BLEU
Random	**46.57**
Perplexity	**46.57**
Formality diff	45.89

Table 8: Translation quality on MSLT-dev when subsampling half number of sentences in OpenSubtitles using different selection strategies. The BLEU score is computed on uncased tokenized segments.

Results in Table 8 show that none of the ranking strategies performs better than ranking randomly. The overlap ratio of selected data using any two methods is slightly larger than 50% – almost by chance. It implied that there were no significant correlations among these data selection method. In addition, comparing with the translation quality of using complete OpenSubtitles (BLEU=47.32), we believe that at least for this specific corpus, useful training examples account for a considerable proportion and that it is challenging to capture the characteristics of Skype conversations with the language model based criteria we considered.

5. Manual Analysis

We[4] manually examine the output of two of our systems on the development data with the goals to (1) determine whether the addition of heuristically-selected training examples had a visible impact on translation quality, and (2) evaluate the quality of the best performing variant of our systems beyond BLEU score and characterize the remaining error patterns.

5.1. Impact of Additional Informal Data

As we have seen in Section 4.6, the impact of additional training data on BLEU score is positive but small. Out of 5262 segments, 1163 were translated differently when incorporating the additional training data. We examine a sample of 100 examples where the two system outputs differ, and find that these differences are small to the human eye as well.

There was no clear impact on adequacy. 9 translations improved with additional training data, and 5 were worse. The two systems produced translation of comparable fluency. In the few instances where there was a noticeable difference, the translations of the augmented system tended to be less fluent (6 examples vs. 4 examples that were more fluent).

The impact on formality is mixed. Out of 100 examples, additional data improved register in 12 examples (mostly due

[4]The second author is a native French speaker.

	Examples
src	got it .
ref	j' ai compris .
mt	je l' ai .
	both reference and MT okay when source segment is out of context
src	savings plans mostly
ref	l' épargne prévoit principalement
mt	des plans d' épargne surtout
	both reference and MT okay when source segment is out of context
src	yeah , that 's how it is , and now that 's what i
ref	ouais , c' est comme a que c' est , et maintenant c' est ce que je
mt	oui , c' est comment c' est , et maintenant c' est ce que j' ai
	formality, disfluency, incomplete source segment
src	no personal details , please .
ref	pas de détails personnels , s' il te plaît .
mt	pas de détails personnels , s' il vous plaît .
	formality difference underspecified in the source segment
src	well , i 've been doing this diet called the paleo diet for about a year now and it has really been great .
ref	ben , je suis en train de faire ce régime appelé régime paleo depuis environ un an maintenant et cela a été vraiment super .
mt	eh bien , j' ai fait ce régime appelé le régime paléo pendant environ un an maintenant , et ca a vraiment été super .
	local disfluencies: different tense, wrong preposition selection, which changes meaning
src	definitely , i totally agree about the more educated you are towards a topic .
ref	sans aucun doute , je suis tout à fait d' accord à propos du sujet sur lequel tu es le plus instruit .
mt	définitivement , je suis totalement d' accord pour les plus instruites tu es vers un sujet .
	local disfluencies: wrong reordering, agreement error, preposition error. Note that gender not specified in English source.
src	but luckily , he has not died yet .
ref	mais heureusement , il n' est pas encore mort .
mt	mais heureusement , il n' a pas encore mort .
	local disfluency: incorrect use of auxiliary "avoir" rather than "être"
src	about touching that money that they saved for retirement
ref	a propos de toucher cet argent qu' ils ont gardé pour la retraite
mt	à propos de toucher à l' argent qu' ils ont sauvé pour la retraite
	incorrect lexical choice

Table 7: Examples of English source segments (src) and their translations (mt) that do not exactly convey the same meaning as the reference (ref). For each example, the nature of the difference between the reference and the MT output is described in italics.

to better selection of informal "tu" over "vous"), and hurt in 9. The augmented system consistently but incorrectly translates "yeah" into "oui" (yes), rather than the more informal "ouais" which is used in reference translations. While this does not alter the meaning of the translation, "ouais" preserves the register of the source more accurately.

5.2. System Evaluation

Out of 100 randomly selected examples, 61 were found to be acceptable translations, that accurately conveyed the meaning of the English segment into French. Among the translations deemed incorrect, the errors were typically localized, and part of the meaning was typically preserved in long sentences. Issues stemmed either from locally disfluent output, lack of grammatical agreement, and more rarely from incorrect lexical choice. Table 7 illustrates the main types of issues observed.

Among examples counted as acceptable, 9 translations had the same meaning as the source but expressed in a more formal way in the MT output than in the reference: e.g., by using the formal second person singular pronoun "vous" rather than the more informal "tu", translating "yeah" as "oui" (yes) rather than the more informal "ouais", or translating "old" as "âgé" rather than "vieux".

There were also 4 examples where MT differed in meaning from the reference, yet were considered acceptable since both MT and reference could be considered correct without discourse context. This is illustrated by the top two examples in Table 7. Translating segments within their discourse context could help address this issue, but the current data release did not provide dialogue boundaries and turns, and our MT system translated segments independently as is typically done in most machine translation architectures.

Fluency and agreement represented the majority of issues with 19 examples. The sixth example in the Table illustrates

some of the phenomena observed: the MT system produced a literal almost word-for-word translation of the English source which resulted in an incorrect preposition ("vers"), a word order that changes the meaning of the source, and the strange use of the feminine plural form in translating "educated" into "instruites". While number should be singular, it should be noted that the gender of the person referred to as "you" is unknown given the source segment alone.

There were 9 cases of disfluent MT due to a segmentation of the source that did not result in a complete sentence (see third example in Table 7). This seems to happen either when the speaker's utterance was not a complete sentence, or when the speaker's utterance was segmented in the transcript.

Lexical choice errors were not a significant issue: there were only 3 cases where lexical choice was considered incorrect, and 5 where the French translation was understandable but sounded awkward. Only two occurrences out-of-vocabulary words were observed.

Overall this analysis suggests that our straightforward system trained on OpenSubtitles data yields surprisingly good translation quality given the lack of in-domain data. OpenSubtitles proves to be an effective substitute, and translations are particularly good for short sentences. Translation errors in longer sentences remain local.

6. Conclusions and Future Work

Our evaluations suggest that a straightforward phrase-based system trained on pseudo in-domain corpus (OpenSubtitles) yields high BLEU scores and errors were mostly local disfluencies. Small improvements in BLEU were obtained by using a simple heuristic to select more Skype-like examples. This approach was the only data selection approach that improved performance. Using language modeling and text formality criterion to select examples that we expected to be closer to Skype data did not improve BLEU.

In future work, we will turn to neural machine translation architectures [16] and will investigate to what extent they can help address the fluency and agreement issues observed here. We are also interested in improving translation selection that is appropriate for the formality level inspired by promising results in preserving politeness in English-German translation [17].

7. References

[1] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, R. Cattoni, and M. Federico, "The IWSLT 2016 Evaluation Campaign," in *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2016)*, Seattle, USA, 2016.

[2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics, 2007.

[3] C. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of IBM model 2," in *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*. The Association for Computational Linguistics, 2013, pp. 644–648.

[4] K. Heafield, "KenLM: faster and smaller language model queries," in *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, United Kingdom, July 2011, pp. 187–197.

[5] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *34th Annual Meeting of the Association for Computational Linguistics, 24-27 June 1996, University of California, Santa Cruz, California, USA, Proceedings.* Morgan Kaufmann Publishers / ACL, 1996, pp. 310–318.

[6] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 7-12 July 2003, Sapporo Convention Center, Sapporo, Japan.* ACL, 2003, pp. 160–167.

[7] K. Yasuda, R. Zhang, H. Yamamoto, and E. Sumita, "Method of selecting training data to build a compact and efficient translation model," in *Third International Joint Conference on Natural Language Processing, IJCNLP 2008, Hyderabad, India, January 7-12, 2008*, 2008, pp. 655–660.

[8] G. F. Foster, C. Goutte, and R. Kuhn, "Discriminative instance weighting for domain adaptation in statistical machine translation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL.* ACL, 2010, pp. 451–459.

[9] R. C. Moore and W. D. Lewis, "Intelligent selection of language model training data," in *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, Short Papers*. The Association for Computer Linguistics, 2010, pp. 220–224.

[10] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *Proceedings*

of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL. ACL, 2011, pp. 355–362.

[11] W. Lewis, C. Federmann, and Y. Xin, "Applying Cross-Entropy Difference for Selecting Parallel Training Data from Publicly Available Sources for Conversational Machine Translation," in *International Workshop on Spoken Language Translation*, 2016.

[12] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: A Resource for the Next Generations of Speech-to-Text." in *LREC*, vol. 4, 2004, pp. 69–71.

[13] F. Heylighen and J.-M. Dewaele, "Variation in the contextuality of language: An empirical measure," *Foundations of Science*, vol. 7, no. 3, pp. 293–340, 2002.

[14] E. Pavlick and J. R. Tetreault, "An empirical analysis of formality in online communication," *TACL*, vol. 4, pp. 61–74, 2016.

[15] E. Pavlick and A. Nenkova, "Inducing lexical style properties for paraphrase and genre differentiation," in *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*. The Association for Computational Linguistics, 2015, pp. 218–224.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[17] R. Sennrich, B. Haddow, and A. Birch, "Controlling Politeness in Neural Machine Translation via Side Constraints." Association for Computational Linguistics, 2016, pp. 35–40.

[18] C. Hardmeier, S. Stymne, J. Tiedemann, and J. Nivre, "Docent: A Document-Level Decoder for Phrase-Based Statistical Machine Translation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 193–198.