Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3775–3781
Marseille, 11–16 May 2020

Corpora for Document-Level Neural Machine Translation

Siyou Liu* **Xiaojun Zhang†**
* Macao Polytechnic Institute
† Xi'an Jiaotong-Liverpool University
violetal@ipm.edu.mo
xiaojun.zhang01@xjtlu.edu.cn

Abstract

Instead of translating sentences in isolation, document-level machine translation aims to capture discourse dependencies across sentences by considering a document as a whole. In recent years, there have been more interests in modelling larger context for the state-of-the-art neural machine translation (NMT). Although various document-level NMT models have shown significant improvements, there nonetheless exist three main problems: 1) compared with sentence-level translation tasks, the data for training robust document-level models are relatively low-resourced; 2) experiments in previous work are conducted on their own datasets which vary in size, domain and language; 3) proposed approaches are implemented on distinct NMT architectures such as recurrent neural networks (RNNs) and self-attention networks (SANs). In this paper, we aim to alleviate the low-resource and under-universality problems for document-level NMT. First, we collect a large number of existing document-level corpora, which covers 7 language pairs and 6 domains. In order to address resource sparsity, we construct a novel document parallel corpus in Chinese–Portuguese, which is a non-English-centred and low-resourced language pair. Besides, we implement and evaluate the commonly-cited document-level method on top of the advanced Transformer model with universal settings. Finally, we not only demonstrate the effectiveness and universality of document-level NMT, but also release the preprocessed data, source code and trained models for comparison and reproducibility.

Keywords: Neural Machine Translation, Document-Level Translation, Corpus, Discourse

1. Introduction

Neural machine translation (NMT) has been rapidly developed in recent years (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015). Conventional NMT models still translate a text by considering isolated sentences, which may harm translation quality especially in terms of coherence, cohesion, and consistency (Webber, 2014). To address this problem, document-level NMT has attracted increasing attention from the community (Wang et al., 2017a; Jean et al., 2017; Kuang et al., 2017b; Bawden et al., 2018; Maruf and Haffari, 2018; Tu et al., 2018; Zhang et al., 2018; Miculicich et al., 2018; Voita et al., 2018; Xiong et al., 2019). Researchers have investigated a variety of approaches to model cross-sentence context for NMT and shown promising results in terms of BLEU score (Papineni et al., 2002) as well as human evaluation (Läubli et al., 2018). For instance, Wang et al. (2017a) tried early attempt proposing a hierarchical neural encoder to summarize context in previous sentences and then integrate the historical representations into standard NMT with three effective strategies.

Although such approaches have achieved significant improvements, there nonetheless exist three main drawbacks that may restrict development of document-level NMT. First, parallel documents are too scarce to build robust document-level NMT models. Taking Chinese–English translation task for example, the WMT2019 sentence-level corpus contains 25 million sentence pairs while the size of IWSLT2017 document-level corpus is only 200 thousands. Second, experiments in previous work are usually conducted on their own datasets which vary in size, domain and language. For example, Wang et al. (2017a) only evaluated their hierarchical model on 1 million Chinese-English LDC corpus in the news domain, while (Jean et al., 2017) mainly verified the multi-encoder method with 200 thousands German-English IWSLT dataset in the spoken domain. Third, different document-level NMT models are implemented on distinct architectures including recurrent neural networks (RNN) (Bahdanau et al., 2015) and self-attention networks (SAN) (Vaswani et al., 2017). Consequently, it is difficult to robustly build document-level models and fairly compare different approaches across different datasets or architectures.

In this paper, we aim to alleviate the low-resource and under-universality problems for document-level NMT. More specifically, we collect a large number of existing corpora for document-level translation task and process them using unified preprocessing steps. As a result, the collected corpora contain 7 language pairs (e.g. Chinese–English, English–French, Estonian–English etc.) in 6 domains (e.g. news, subtitle, parliament etc.). In addition, we construct a new document-level corpus in Chinese–Portuguese, which is low-resourced and non-English-centred language pair. Finally, we implement a commonly-used document-aware approach (Wang et al., 2017a; Jean et al., 2017; Voita et al., 2018) on top of a state-of-the-art SAN-based NMT model – Transformer.

Experiments are systematically conducted on all collected and built data using Transformer models. Results confirm that the document-level information is indeed useful to NMT. For better comparison and reproducibility, we release the preprocessed data,[1] source code and trained models. We hope this work can be used as a benchmark for other researchers to further improve document-level NMT. The contributions of this paper are listed as follows:

- We investigate a variety of the document-level corpora for NMT, which confirms the superiority of modelling

[1] For some non-public corpora such as LDC, we provide the download links and preprocessing scripts.

larger context;

- We build a novel document-level parallel corpus for low-resourced language pair, and experiments show that document context is more helpful to long-distant languages.
- For better comparison and reproducibility, we release the preprocessed data, source code and trained models.

The rest of the paper is organized as follows. In Section 2, we introduce background of NMT as well as the document-level NMT model. The details of document-level parallel corpora are described in Section 3. The experimental results of translation task using different data are reported in Section 4. Related work are given in Section 5. Finally, Section 6 presents our conclusions and future work.

2. Background

2.1. Neural Machine Translation

A standard NMT model directly optimizes the conditional probability of a target sentence $\mathbf{y} = y_1, \ldots, y_J$ given its corresponding source sentence $\mathbf{x} = x_1, \ldots, x_I$:

$$P(\mathbf{y}|\mathbf{x};\theta) = \prod_{j=1}^{J} P(y_j|\mathbf{y}_{<j},\mathbf{x};\theta) \qquad (1)$$

where θ is a set of model parameters and $\mathbf{y}_{<j}$ denotes the partial translation. The probability $P(\mathbf{y}|\mathbf{x};\theta)$ is defined on the neural network based encoder-decoder framework (Sutskever et al., 2014; Cho et al., 2014), where the encoder summarizes the source sentence into a sequence of representations $\mathbf{H} = \mathbf{H}_1, \ldots, \mathbf{H}_I$ with $\mathbf{H} \in \mathbb{R}^{I \times d}$, and the decoder generates target words based on the representations. Typically, this framework can be implemented as recurrent neural network (RNN) (Bahdanau et al., 2015), convolutional neural network (CNN) (Gehring et al., 2017) and Transformer (Vaswani et al., 2017). Among the different models, the Transformer has emerged as the dominant NMT paradigm. In this study, we re-implement the baseline and document-level models on top of Transformer.
The parameters of the NMT model are trained to maximize the likelihood of a set of training examples $D = \{[\mathbf{x}^m, \mathbf{y}^m]\}_{m=1}^{M}$:

$$\mathcal{L}(\theta) = \arg\max_{\theta} \sum_{m=1}^{M} \log P(\mathbf{y}^m|\mathbf{x}^m;\theta) \qquad (2)$$

which is used as a sentence-level baseline in this work.

2.2. Motivation

As shown in Section 2.1, the standard NMT usually models a text by considering isolated sentences based on a strict assumption that the sentences in a text are independent of one another. However, disregarding dependencies across sentences will negatively affect translation outputs of a text in terms of discourse properties.
Coherence, cohesion, and consistency are three main properties of discourse. *Cohesion* occurs whenever "the interpretation of some element in the discourse is dependent on that of another" (Halliday and Hasan, 1976), which refers to various manifest linguistic links (e.g. references, word repetitions) between sentences within a text that holds the text together. Coherence is created referentially, when different parts of a text refer to the same entities, and relationally, by means of coherence relations such as "Cause–Consequence" between different discourse segments. Consistency is another critical issue in document-level translation, where a repeated term should keep the same translation throughout the whole document. The underlying assumption is that the same concepts should be consistently referred to with the same words in a translation.
Recent studies have shown that incorporating document-level is helpful to translations in terms of coherence (Wang et al., 2016b; Xiong et al., 2019), cohesion (Wang et al., 2016a; Voita et al., 2018; Wang et al., 2018b; Wang et al., 2018a), and consistency (Xiao et al., 2011; Wang et al., 2017a; Wang et al., 2019). This motivated us to continue exploit document-level NMT.

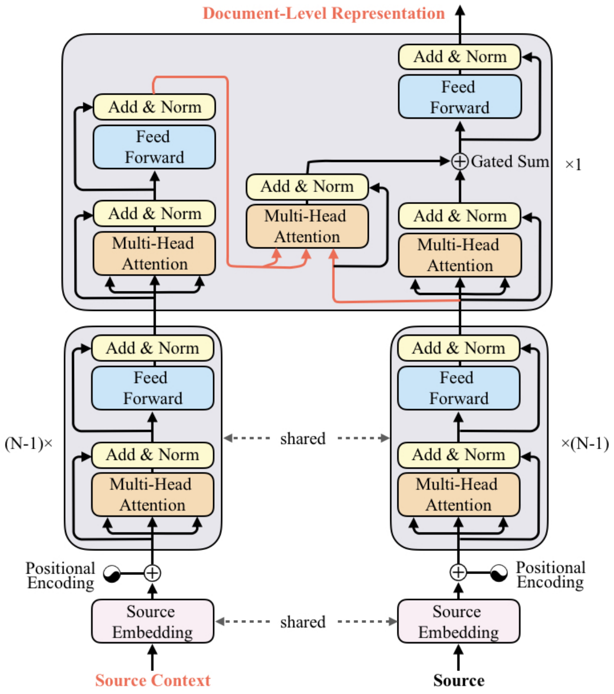

Figure 1: The architecture of a multi-encoder model. As seen, the encoder is composed of a stack of N layers, each of which contains self-attention and feed-forward sublayers. "Source" and "Source Context" represent current sentence and its previous sentences in a document.

2.3. Document-Level Neural Machine Translation

This task aims to consider both the current sentence and its large context in a unified model to improve translation performances, especially in terms of discourse properties. Wang et al. (2017a), Jean et al. (2017) and Voita et al. (2018) proposed a novel document-level NMT model, namely *multi-encoder*.
As shown in Figure 1, they employed $(N-1)\times$ layers of context encoder to summarize the larger context from

source-side previous sentences, and $(N-1)\times$ layers of standard encoder to model the current sentence. At the last layer, they integrate the contextual information with the source representations using a gating mechanism. Finally, the combined document-level representations are fed into NMT decoder to translate the current sentence.

Given a source sentence $\mathbf{x}_i$ to be translated, we consider its K previous sentences in the same document as source context $C = \{\mathbf{x}_{i-K}, ..., \mathbf{x}_{i-1}\}$. The source encoder employs multi-head self-attention $\textsc{Att}(\cdot)$ to transform an input sentence $\mathbf{x}_i$ into a sequence of representations $\mathbf{O}^h = \{\mathbf{o}_1^h, \dots, \mathbf{o}_I^h\}$ by:

$$\mathbf{o}_i^h = \textsc{Att}(\mathbf{q}_i^h, \mathbf{K}^h)\mathbf{V}^h \in \mathbb{R}^{\frac{d}{H}} \tag{3}$$

where h is one of H heads. $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ respectively represent queries, keys and values, which are calculated as:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{X}\mathbf{W}_Q, \mathbf{X}\mathbf{W}_K, \mathbf{X}\mathbf{W}_V \in \mathbb{R}^{I \times d} \tag{4}$$

where $\{\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V\} \in \mathbb{R}^{d\times d}$ are trainable parameters and d indicates the hidden size. The context encoder employs the same networks as source encoder to obtain the context output $\hat{\mathbf{O}}$. Finally, the two encoder outputs $\mathbf{O}$ and $\hat{\mathbf{O}}$ are combined via a gated sum, as in:

$$\lambda_d = \sigma(W_\lambda[O_d, \hat{O}_d] + b_d) \tag{5}$$

$$O' = \lambda_d \odot O_d + (1-\lambda_d) \odot \hat{O}_d \tag{6}$$

in which $\sigma(\cdot)$ is the logistic sigmoid function and W_λ is the parameter. O' is the final document-level representation, which is further fed into NMT decoder. Following Voita et al. (2018), we share the parameters of context encoders and embedding with those of standard NMT encoder.

3. Document-Level Parallel Corpora

We reviewed a large number related work on document-level NMT (Wang et al., 2017a; Jean et al., 2017; Kuang et al., 2017b; Bawden et al., 2018; Maruf and Haffari, 2018; Tu et al., 2018; Zhang et al., 2018; Miculicich et al., 2018; Voita et al., 2018; Xiong et al., 2019), and collected existing data used in their experiments. In addition, we also built a new document-level Chinese–Portuguese corpus based on our previous work Liu et al. (2018).

Table 1 lists all document-level parallel corpora, each of which differs from others in language, topic, genre, style, level of formality, etc.

3.1. Collected Corpora

LDC Most sentences in this corpus come from the news and law domains. They are formal articles with syntactic structures such as complicated conjoined phrases, which make textual translation very difficult. The sub-corpora indexes of training data are: 2003E07, 2003E14, 2004T07, 2005E83, 2005T06, 2006E24, 2006E34, 2006E85, 2006E92, 2007E87, 2007E101, 2007T09, 2008E40, 2008E56, 2009E16, 2009E95, 2005T10, 2004T08, 2002T01, 2009T02, 2009T15, 2010T03. The NIST 2002 (MT02) dataset is used as tuning set, and the NIST 2003-2008 (MT03-MT08) datasets as test sets. It only have Chinese–English language pair.[2]

OpenSubtitle This is a collection of translated movie subtitles (Lison and Tiedemann, 2016), which are usually simple and short. Most of the translations of subtitles do not preserve the syntactic structures of their original sentences at all. We randomly select two episodes as the tuning set, and the other two episodes as the test set. It totally contains 62 language pairs and here we mainly exploited commonly-cited French–English, Spanish–English and Russian–English.[3]

IWSLT The corpora are from the machine translation track on TED Talks of IWSLT (Cettolo et al., 2012). Koehn and Knowles (2017) point out that NMT systems have a steeper learning curve with respect to the amount of training data, resulting in worse quality in low-resource settings. The TED talks are difficult to translate given the variety of topics in quite small-scale training data. We choose the "dev2010" dataset as the tuning set, and the combination of "tst2010-2013" datasets as the test set. It totally contains nearly 100 language pairs while here we mainly investigated commonly-used Chinese-English, French-English, Spanish–English and German–English.[4]

News Commentary The corpus was created as training data resource for the Conference for Statistical Machine Translation Evaluation Campaign and consists of political and economic commentary crawled from the web site Project Syndicate (Koehn, 2005). Different from LDC, this corpus mainly focuses on political/economic news with more language pairs. It totally contains 12 language pairs and here we evaluated Spanish–English and German–English.[5]

Europarl The corpus is extracted from the proceedings of the European Parliament. Sentences are usually long and formal used in the official conference. It contains 21 European language pairs and we used Estonian–English dataset.[6]

TVSub Wang et al. (2018a) extracted subtitles from TV episodes, instead of movies compared with the OpenSubtitle Corpus. Thus, utterances in this dataset is more discourse-aware and it only has Chinese-English language pair.[7]

3.2. Our Corpus

This subsection describes our new document-level parallel corpus in Chinese–Portuguese. The bilingual data were originally extracted from Macao government websites[8] by Liu et al. (2018). Its domain contains international communication, trade exchanges, technological cooperation, etc. In order to build a high-quality parallel corpus, we propose a hierarchical strategy to deal with document-level, paragraph-level and sentence-level alignment. As shown in Table 1, more than 800 thousands of Chinese–Portuguese

[2] https://www.ldc.upenn.edu.

[3] http://opus.nlpl.eu/OpenSubtitles2018.php.

[4] https://wit3.fbk.eu.

[5] http://www.casmacat.eu/corpus/news-commentary.html.

[6] https://www.statmt.org/europarl.

[7] https://github.com/longyuewangdcu/tvsub.

[8] https://news.gov.mo.

Corpus	Language	Domain	Size			Related Work
			$\|S\|$	$\|D\|$	$\|L\|$	
Existing corpora in related work						
LDC	ZH-EN (small)	news	1.3M	22K	22.3/27.6	Wang et al. (2017b); Tu et al. (2018); Zhang et al. (2018)
	ZH-EN (large)	news, law	2.8M	61K	23.7/29.2	Kuang et al. (2017b)
OpenSub	FR-EN	subtitle	29.2M	35K	8.0/7.5	Bawden et al. (2018)
	ES-EN	subtitle	64.7M	78K	8.0/7.3	Miculicich et al. (2018)
	EN-RU	subtitle	27.4M	35K	5.8/6.7	Voita et al. (2018)
IWSLT	ZH-EN	spoken	0.2M	2K	19.5/21.0	Xiong et al. (2019); Tu et al. (2018); Miculicich et al. (2018)
	FR-EN (small)	spoken	0.1M	1K	20.3/21.1	Maruf and Haffari (2018)
	FR-EN (large)	spoken	0.2M	2K	20.8/21.9	Zhang et al. (2018)
	ES-EN	spoken	0.2M	2K	19.9/20.8	Miculicich et al. (2018)
	DE-EN	spoken	0.2M	2K	19.3/20.7	Jean et al. (2017)
NewsCom	ES-EN	news	0.2M	5K	30.7/31.2	Miculicich et al. (2018)
	DE-EN	news	0.2M	5K	33.1/32.2	Maruf and Haffari (2018)
Europarl	ET-EN	parliament	0.2M	150K	35.1/36.4	Maruf and Haffari (2018)
TVsub	ZH-EN	subtitle	2.2M	3K	5.6/7.7	Miculicich et al. (2018); Tu et al. (2018)
New corpus in this work						
MacaoGov	ZH-PT	news, law	0.8M	5K	20.4/26.9	Liu et al. (2018)

Table 1: Statistics of training corpora for document-level NMT. The details are name of corpora, language pairs, domains, number of sentences ($|S|$), number of documents ($|D|$), averaged sentence length ($|L|$) and related work. K stands for thousand and M for million.

sentence pairs in newswire, law and travelling domains, among others, are curated. As Chinese and Portuguese are long distant languages, it is more interesting to see whether document-level context is still useful under such scenario.

3.3. Preprocessing

To preprocess the raw data, we use a series of scripts including: full/half-width conversion, Unicode conversion, simplified/traditional Chinese conversion, punctuation normalization, tokenization and sentence boundary detection, letter casing and word stemming (Wang et al., 2016b). To the end, we employ these methods to uniformly preprocess data described in Section 3.1 and 3.2. Note that, we keep the contextual information for document-level tasks while use only single sentences for sentence-level tasks.

4. Experiment

4.1. Setup

For fair comparison, we implemented baseline and document-level NMT model on the advanced *Transformer* model (Vaswani et al., 2017) using the open-source toolkit Fairseq (Ott et al., 2019). We followed Vaswani et al. (2017) to set the configurations of the NMT model, which consists of 6 stacked encoder/decoder layers with the layer size being 512. All the models were trained on 8 NVIDIA P40 GPUs where each was allocated with a batch size of 4,096 tokens. We trained the baseline model for 100K updates using Adam optimizer (Kingma and Ba, 2015), and the proposed models were further trained with corresponding parameters initialized by the pre-trained baseline model. We fixed the hyperparameters λ and δ as 0.1. According to previous studies (Wang et al., 2017a; Tu et al., 2018), we modeled previous $K = 3$ sentences as document contexts for each current sentence. For the additional encoder, we use the same settings with the standard one as introduced in Figure 1.

We trained standard NMT models on sentence-level data as our baselines (*Base*) and built document-level NMT models ("*DNMT*") using parallel documents as discussed in Section 3.3. Furthermore, we used case-insensitive 4-gram NIST BLEU metrics (Papineni et al., 2002) for evaluation, and *sign-test* (Collins et al., 2005) to test for statistical significance.

4.2. Results

Table 2 shows translation results on different data as described in Section 3. The multi-encoder model ("DNMT") is trained and evaluated on document-level data while the baseline model is trained and evaluated on the corresponding data, which are broken into sentence level. As seen,

Corpus	Language	BLEU		△
		Base	DNMT	
Existing corpora in related work				
LDC	ZH-EN (s)	35.73	36.43†	+0.70
	ZH-EN (l)	34.21	35.18†	+0.97
OpenSub	FR-EN	23.64	24.20†	+0.56
	ES-EN	35.25	35.46	+0.21
	EN-RU	29.54	30.16†	+0.62
IWSLT	ZH-EN	18.50	18.79	+0.29
	FR-EN (s)	21.36	22.66†	+1.30
	FR-EN (l)	35.11	35.85†	+0.74
	ES-EN	35.19	35.88†	+0.69
	DE-EN	20.11	20.81†	+0.70
NewsCom	ES-EN	21.36	22.36†	+1.00
	DE-EN	9.22	10.13†	+0.91
Europarl	ET-EN	20.75	22.41†	+1.66
TVsub	ZH-EN	33.04	33.42†	+0.38
New corpus in this work				
MacaoGov	ZH-PT	26.06	27.09†	+1.03

Table 2: Translation results on a variety of corpora. "Base" is the baseline model trained and evaluated on the sentence-level data, while "DNMT" is multi-encoder model trained and evaluated on corresponding document-level data. "†" indicates statistically significant difference ($p < 0.01$) from "Transformer" in the corresponding corpora.

the document-level models significantly improve the translation quality in all cases, although there are considerable differences among different scenarios. Experimental results confirm that the document-level information indeed improves translation performances on various corpora.

In formal domain of corpora such as LDC, News Commentary, Europarl and MacaoGov, the document-level model achieves larger improvements over the Transformer baseline (+0.7 $\sim$ +1.7 BLEU points). Taking news domain for example, one entity word usually needs to keep consistent translation across the whole document in newswire. Thus, the gains mainly come from better translation consistency contributed by document context.

However, in informal domain such as OpenSubtitle and TV-Sub, the improvements are relatively smaller (+0.2 $\sim$ +0.6 BLEU point) compared with formal domains. As known that, informal data such as dialogue are often difficult to translate due to a lot of discourse phenomena such zero anaphora (Wang et al., 2016a). Another reason maybe that $K = 3$ previous sentences are not enough to recall missing information in the context.

In our experiments, we also evaluated their performances using different matrices, including METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2009). We found that the trends are simlar to BLEU scores.

5. Related Work

5.1. On Document-Level Translation

In early studies, document-aware approaches have been investigated for statistical machine translation (SMT) (Tiedemann, 2010; Gong et al., 2011; Xiao et al., 2011; Hardmeier et al., 2012).

In recent years, context-aware architecture has been well studied for NMT (Wang et al., 2017a; Jean et al., 2017; Tu et al., 2018). Wang et al. (2017a) proposed hierarchical recurrent neural networks to summarize inter-sentential context from previous sentences and then integrate it into a standard NMT model with difference strategies. Jean et al. (2017) introduced an additional set of an encoder and attention to encode and select part of the previous source sentence for generating each target word. Besides, Tu et al. (2018) proposed to augment NMT models with a cache-like memory network, which stores the translation history in terms of bilingual hidden representations at decoding steps of previous sentences. They also evaluated the above three models on different domains of data, showing that the hierarchical encoder performs comparable with the multi-attention model.

5.2. On Discourse Phenomena

More recently, some researchers began to investigate the effects of context-aware NMT on cross-lingual pronoun prediction (Bawden et al., 2018; Voita et al., 2018; Jean and Cho, 2019). They mainly exploited general anaphora in non-pro-drop languages such as English⇒Russian.

In order to evaluate discourse phenomena, Bawden et al. (2018) conducted experiments from three aspects: 1) comparing multi-encoder models (Zoph and Knight, 2016; Jean et al., 2017) with different strategies; 2) investigating the impacts of source- and target-side history information on NMT; 3) presenting a novel evaluation through the use of two discourse test sets targeted at coreference and lexical coherence/cohesion. Voita et al. (2018) introduced a context-aware model and demonstrated its usefulness for anaphora resolution as well as translation. Besides, Xiong et al. (2019) proposed to use discourse context and reward to refine the translation quality from the perspective of coherence. Some researchers proposed to extend the Transformer model to take advantage of document-level context (Miculicich et al., 2018; Zhang et al., 2018). Following Tu et al. (2018)'s work, Kuang et al. (2017a) and Maruf and Haffari (2018) continue to exploit cache memory for improving the performance of document-level NMT. Through human evaluation, Läubli et al. (2018) found that document-level evaluation for MT can improve to discriminate the errors which are hard or impossible to spot at the sentence level.

6. Conclusion and Future Work

In this paper we collected and preprocessed a large number of corpora for document-level translation task. Besides, we implemented and evaluated the document-aware approach on top of a universal NMT model – Transformer. We also construct an additional corpus in a novel language pair (Chinese–Portuguese). We conduct experiments on existing and the curated corpora, and compare the performance of different NMT models using these corpora. Results showed that document contexts are more useful to formal domains than informal ones. We hope this work can be used by MT research for further improving document-level translation.

In the future, we will investigate more document-level approaches such as the hierarchical encoder proposed by Wang et al. (2017a; Miculicich et al. (2018). Furthermore, we will continue to exploit Zhang et al. (2018)'s training strategy to make full use of large-amount sentence-level data.

7. Acknowledgements

This work is supported by the XJTLU KSF project (Grant Number: KSF-E-24) and GDUFS open project (Grant Number:CTS201501). The authors also wish to thank the anonymous reviewers for many helpful comments.

8. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *ICLR*.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In *NAACL*.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. In *EAMT*.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*.

Collins, M., Koehn, P., and Kucerova, I. (2005). Clause restructuring for statistical machine translation. In *ACL*.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *ICML*.

Gong, Z., Zhang, M., and Zhou, G. (2011). Cache-based document-level statistical machine translation. In *EMNLP*.

Halliday, M. A. K. and Hasan, R. (1976). Cohesion in english. *Longman*.

Hardmeier, C., Nivre, J., and Tiedemann, J. (2012). Document-wide decoding for phrase-based statistical machine translation. In *EMNLP*.

Jean, S. and Cho, K. (2019). Context-aware learning for neural machine translation. *arXiv preprint arXiv:1903.04715*.

Jean, S., Lauly, S., Firat, O., and Cho, K. (2017). Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *EMNLP*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*.

Kuang, S., Xiong, D., Luo, W., and Zhou, G. (2017a). Cache-based document-level neural machine translation. *arXiv preprint arXiv:1711.11221*.

Kuang, S., Xiong, D., Luo, W., and Zhou, G. (2017b). Modeling coherence for neural machine translation with dynamic and topic caches. *arXiv preprint arXiv:1711.11221*.

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? a case for document-level evaluation. In *EMNLP*.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *LREC*.

Liu, S., Wang, L., and Liu, C.-H. (2018). Chinese-portuguese machine translation: A study on building parallel corpora from comparable texts. In *LREC*.

Maruf, S. and Haffari, G. (2018). Document context neural machine translation with memory networks. In *ACL*.

Miculicich, L., Ram, D., Pappas, N., and Henderson, J. (2018). Document-level neural machine translation with hierarchical attention networks. In *EMNLP*.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). Fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *ACL*.

Snover, M. G., Madnani, N., Dorr, B., and Schwartz, R. (2009). Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23(2-3):117–127.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS*.

Tiedemann, J. (2010). Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*.

Tu, Z., Liu, Y., Shi, S., and Zhang, T. (2018). Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*.

Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (2018). Context-aware neural machine translation learns anaphora resolution. In *ACL*.

Wang, L., Tu, Z., Zhang, X., Li, H., Way, A., and Liu, Q. (2016a). A novel approach to dropped pronoun translation. In *NAACL*.

Wang, L., Zhang, X., Tu, Z., Way, A., and Liu, Q. (2016b).

Automatic construction of discourse corpora for dialogue translation. In *LREC*.

Wang, L., Tu, Z., Way, A., and Liu, Q. (2017a). Exploiting cross-sentence context for neural machine translation. In *EMNLP*.

Wang, R., Finch, A., Utiyama, M., and Sumita, E. (2017b). Sentence embedding for neural machine translation domain adaptation. In *ACL*.

Wang, L., Tu, Z., Shi, S., Zhang, T., Graham, Y., and Liu, Q. (2018a). Translating pro-drop languages with reconstruction models. In *AAAI*.

Wang, L., Tu, Z., Way, A., and Liu, Q. (2018b). Learning to jointly translate and predict dropped pronouns with a shared reconstruction mechanism. In *EMNLP*.

Wang, L., Tu, Z., Wang, X., and Shi, S. (2019). One model to learn both: Zero pronoun prediction and translation. In *EMNLP*.

Webber, B. (2014). Discourse for machine translation. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*.

Xiao, T., Zhu, J., Yao, S., and Zhang, H. (2011). Document-level consistency verification in machine translation. In *MT Summit*.

Xiong, H., He, Z., Wu, H., and Wang, H. (2019). Modeling coherence for discourse neural machine translation. In *AAAI*.

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. (2018). Improving the transformer translation model with document-level context. In *EMNLP*.

Zoph, B. and Knight, K. (2016). Multi-source neural translation. In *NAACL*.