A Test Suite for Evaluating Discourse Phenomena in Document-level Neural Machine Translation

Xinyi Cai
College of Intelligence and Computing,
Tianjin University, China
xinyicai@tju.edu.cn

Deyi Xiong
College of Intelligence and Computing,
Tianjin University, China
dyxiong@tju.edu.cn

Abstract

The need to evaluate the ability of context-aware neural machine translation (NMT) models in dealing with specific discourse phenomena arises in document-level NMT. However, test sets that satisfy this need are rare. In this paper, we propose a test suite to evaluate three common discourse phenomena in English-Chinese translation: pronoun, discourse connective and ellipsis where discourse divergences lie across the two languages. The test suite contains 1,200 instances, 400 for each type of discourse phenomena. We perform both automatic and human evaluation with three state-of-the-art context-aware NMT models on the proposed test suite. Results suggest that our test suite can be used as a challenging benchmark test bed for evaluating document-level NMT. The test suite will be publicly available soon.

1 Introduction

Document-level NMT has attracted extensive interest in recent years. Different from sentence-level NMT models, discourse-level models need to not only cope with intra-sentence dependencies, but also incorporate context beyond current sentence into context-aware translation. Inter-sentence links usually exhibit a wide variety of discourse phenomena: coreference, lexical cohesion, coherence, discourse relations, etc. The quality of a document-level NMT model therefore can be evaluated based on its ability in dealing with these discourse phenomena.

Widely-used automatic evaluation metrics, e.g., BLEU (Papineni et al., 2002), normally consider fragments in a local window for translation quality assessment, while cross-sentence discourse links are usually neglected. Hence, for document-level models, current automatic evaluation metrics may be not a reasonably good fit for evaluation. One possible alternative is using manually-created test suites which are composed of carefully selected examples with discourse phenomena (Hardmeier, 2015).

Such test suites (Guillou et al., 2018; Rysová et al., 2019; Vojtěchová et al., 2019; Voita et al., 2019; Popović, 2019) have been constructed for several language pairs, such as English-Czech, English-German, English-Russian, French-German, but few in English-Chinese translation. In this paper, we propose a test suite aiming at English-Chinese discourse phenomena evaluation. Three frequent discourse phenomena in English-Chinese translation are selected in our test suite, namely pronoun, discourse connective and ellipsis, each of which forms an individual test set. We choose examples from the OpenSubtitles (Lison and Tiedemann, 2016) to construct the three test sets. Unlike corpora from news domain, this corpus is more conversational and colloquial. We use this test suite to evaluate several typical context-aware NMT models. The experiment results show that our test suite can evaluate the ability of NMT models in dealing with discourse phenomena and that it is still very challenging for current context-aware models to capture different discourse phenomena.

2 Related Work

Research on the evaluation of document-level machine translation is usually on specific discourse phenomena. A few test suites and methods have been designed for evaluating NMT from the perspective of discourse phenomena.

For **pronoun** translation evaluation, recent test sets on pronoun evaluation have consisted of contrastive pairs. Bawden et al. (2018) provide 50 example blocks of English-French contrastive pairs. Müller et al. (2018) have also created contrastive pairs of pronoun “it” in English-German translation.

Proceedings of the 2nd International Workshop on Discourse Processing, pages 13–17
Suzhou, China, December 7, 2020. ©2020 Association for Computational Linguistics

Contrastive test sets allow us to automatically evaluate document-level NMT by only judging whether the evaluated model can choose the correct translation against the wrong from each contrastive pair according to their model score. However, this is an indirect rather than a direct way to evaluate the ability of context-aware NMT in modeling discourse phenomena as we do not evaluate the actual translations generated by these NMT systems.

To evaluate **discourse connective** translation, Meyer et al. (2012) propose ACT (accuracy of connective translation) to evaluate connective translation. For French-English discourse relation and discourse connective translation assessment, Smith and Specia (2018) use pretrained bilingual embeddings of discourse connectives. Popović (2019) investigates conjunction disambiguation in English-German and French-German translation.

For the evaluation on **ellipsis** translation, Voita et al. (2019) explore contrastive examples to evaluate the verb phrase ellipsis and morphological inflection in English-Russian translation. In our work, we also investigate verb ellipsis in English-Chinese translation.

3 Test Sets

We choose three types of discourse phenomena, i.e., pronoun, discourse connective and ellipsis, as they appear frequently in English-Chinese document-level NMT. In the following parts, we will introduce corpus construction and then the three test sets separately.

3.1 Test Sets Construction

Due to the lack of such a test set for English-Chinese translation, we manually construct our test sets. We select instances from the open-source corpus OpenSubtitles (Lison and Tiedemann, 2016) as our data sources. First, we filter out characters and tokens written in languages other than English and Chinese. We then extract snippets with two neighboring sentences. Finally, we select test cases from extracted snippets according to different language phenomena.

For the construction of the pronoun test set, we discard snippets where the two adjacent sentences both include "you" or "they" in English. We then construct the test set from the remaining examples that contain "你", "你们", "她们", "它们" and "他们" on the Chinese side.

For the construction of the discourse connective test set, we automatically select examples where the second sentence contains specific discourse connectives. From these examples we manually select samples where English sentences contain ambiguous connectives with different senses, according to Webber et al. (2019).

source:
context: **You** rich guys think that money can buy anything.
current: How right **you** are.
target:
context: 你们富人总以为钱能买到一切。
current: 你们想的太对了。

Figure 1: An example from the pronoun test set.

As for the ellipsis test set, we first choose cases where the second sentence in English contains auxiliary verbs. If the Chinese translations of the chosen cases whether include ellipsis verbs, such cases are finally selected.

As Chinese translations are provided by non-professional translators, they are sometimes noisy with errors. We hire professional translators to review the selected instances and correct translation errors. Each test set contains 400 examples. Data statistics are displayed in Table 1.

3.2 Pronoun Test Set

In the pronoun test set, we focus on the second person pronoun "you" and the third person pronoun "they" as well as their accusative and possessive forms. In Chinese, "you" can be translated as "你" (single form) or "你们" (plural form). And "they" is translated into words of different genders: "他们" (plural form of "he"), "她们" (plural form of "she") and "它们" (plural form of "it"). Each type of pronouns has 80 examples in this test set. Figure 1 displays an example from this test set.

In order to help document-level NMT models choose a correct translation for "you" and "they", we provide the previous sentence as context, which is guaranteed to elliminate such translation ambiguity. For "you", the preceding sentence usually contains nouns or names which indicate the plural or single information of the pronoun. As for "they", nouns with gender information, common names of men and women or non-human nouns in the source side context can be explored for translation disambiguation.

3.3 Discourse Connective Test Set

For this test set, we focus on ambiguous discourse connective in English-Chinese translation. Particu-

	sentences	words		words/sentence		discourse phenomena
		en	zh	en	zh	
pronoun	800	8,469	7,871	10.59	9.84	400
connective	800	10,398	9,691	13	12.11	400
ellipsis	800	5,913	5,630	7.39	7.04	400

Table 1: Data statistics.

source:
context: Everything is so difficult in life, for me.
current: **While** for others it's all child's play.
target:
context: 对于我，生活一切都很艰难。
current: 对于别人却都像儿戏一样。

Figure 2: An example from the discourse connective test set.

source:
context: You see, she doesn't **know**.
current: Neither do I.
target:
context: 看，她不知道。
current: 我也不知道。

Figure 3: An example from the ellipsis test set.

larly, we select five ambiguous discourse connectives according to Webber et al. (2019), namely *while*, *as*, *since*, *though* and *or*. Different senses of these ambiguous connectives are frequently occurring in English texts. The number of cases for each connective is 80. An example of discourse connective in this test set is demonstrated in Figure 2.

Discourse connectives are important to express the discourse relation between sentences. The same connective in different context, may convey different discourse relations in the sense hierarchy (Webber et al., 2019). In order to correctly translate these ambiguous connectives, context-aware NMT models have to recognize discourse relations between clauses or sentences by taking sufficient context into account.

3.4 Ellipsis Test Set

We cover verb ellipsis in English in this test set. As illustrated in Figure 3, Chinese and English exhibit different ellipsis patterns, which pose challenges for machine translation.

If we are only given a sentence with ellipsis, we cannot fully understand this sentence as crucial information may be missing, which can only be recovered by resorting to previous context. For context-aware NMT models, this means that they have to find the elided information if this information should be present in the target language.

4 Experiment

We used the proposed test suite as a benchmark test bed to evaluate state-of-the-art context-aware NMT models against the three types of discourse phenomena.

4.1 Models

We used the following three document-level NMT models:

- *thumt*: Zhang et al. (2018) extend the Transformer model with a new context encoder to model document-level context, which is then incorporated into the original encoder and decoder. They introduce a two-step training method to explore abundant sentence-level parallel corpora and limited document-level parallel corpora.
- *CADec*: Voita et al. (2019) introduce a two-pass framework, which first translates a sentence with a context-agnostic model and refines the target translation with both source and target context.
- *bert-nmt*: Zhu et al. (2020) propose a BERT-fused model. They first use BERT to extract representations for an input sequence, and then fuse the representations into each layer of the encoder and decoder of the NMT model through attention mechanisms.

4.2 Data

We used the following corpora to train the three NMT models: 6M sentence pairs randomly selected from AI Challenger[1] 2017 English-Chinese machine translation corpus, IWSLT'17 training data and a subset of OpenSubtitles (Lison and Tiedemann, 2016) with 1.27M sentence pairs, where instances in our test suite were excluded. The AI Challenger 2017 is a sentence-level MT

[1] https://challenger.ai/

	pronoun	connective	ellipsis
thumt	12.4	9.8	18.2
CADec	**19.1**	**15.3**	**25.5**
bert-nmt	13.9	12.7	19.1

Table 2: BLEU scores on the three test sets.

	you (pl.)	you (sing.)	it (pl.)	she (pl.)	he (pl.)	while	as	since	though	or	pronoun	connective	ellipsis
thumt	8.75	**97.5**	11.25	7.5	91.25	**57.5**	30	41.25	48.75	67.5	49	43.25	**10.75**
CADec	**22.5**	96.25	33.75	**20**	**92.5**	48.75	**53.75**	**56.25**	**57.5**	**83.75**	**53**	**60**	2.25
bert-nmt	18.75	93.75	**38.75**	0	90	53.75	31.25	41.25	46.25	75	48.25	49.5	5.75

Table 3: Human evaluation results (accuracy %) on the three test sets.

corpus in spoken language. IWSLT'17 English-Chinese MT corpus comprises of TED talks.

Thumt and *CADec* were trained on the sentence-level data, i.e., the 6M-sentence subset of the AI Challenger 2017 corpus, in the first stage. In the second phase of context-aware training, the combination of the IWSLT'17 training data and the subset of the OpenSubtitles corpus was used. *Bert-nmt* was trained on only IWSLT'17 data following Zhu et al. (2020).

4.3 Results

The BLEU scores of the three models on our test suite are shown in Table 2. In addition to the automatic evaluation, we further performed human evaluation to investigate the translation accuracy on the three types of discourse phenomena. In human evaluation, we focus on whether the relevant phenomena are correctly translated and ignore other errors. Human evaluation is better at evaluating discourse phenomena translation. Human evaluation results are shown in Table 3.

Overall, *CADec* achieves the best results in most cases but not in all cases. In translating *you (sing.)*, *while* and *ellipsis*, *thumt* achieves the highest accuracy, while *bert-nmt* is better than the others in translating *they (it (pl.))*.

For pronoun translation, "you" is usually translated into "你" (you (sing.)) while "they" into "他们" (he (pl.)). This is because these two cases are more frequent than other cases (e.g., "你们", "它们"). This also happens for discourse connective translation. For example, "while" is often translated into "当……时候" rather than "而" (but) as the former is more common that the latter.

Compared with pronouns and discourse connectives, ellipsis is more challenging for the three context-aware models, which achieves a translation accuracy of <11%. Verb ellipsis usually occurs in questions or replies in spoken dialogues. We observe that auxiliary verb "do" is often wrongly translated into "do" (notional verb) or "know". This suggests that these context-aware models cannot correctly recognize ellipsis and detect omitted fragments from context.

5 Conclusion

We have presented a discourse-level test suite for the evaluation of context-aware neural machine translation. We constructed 1,200 instances for three types of discourse phenomena in English-Chinese translation, 400 instances per discourse phenomenon. Our experiments with three state-of-the-art document-level NMT models suggest that ellipsis is the most challenging discourse issue among the three test sets.

Acknowledgments

The present research was supported by the National Natural Science Foundation of China (Grant No. 61861130364), Natural Science Foundation of Tianjin (Grant No.19JCZDJC31400) and a Newton International Fellowship from the Royal Society (London)(NAF\R1\180122). We would like to thank the anonymous reviewers for their insightful comments. The corresponding author is Professor Deyi Xiong (dyxiong@tju.edu.cn).

References

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018.

A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Christian Hardmeier. 2015. On statistical machine translation and translation theory. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 168–172, Lisbon, Portugal. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *LREC*.

Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. 2012. Machine translation of labeled discourse connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, page 10.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. 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 on Association for Computational Linguistics*, ACL ’ 02, page 311318, USA. Association for Computational Linguistics.

Maja Popović. 2019. Evaluating conjunction disambiguation on English-to-German and French-to-German WMT 2019 translation hypotheses. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 464–469, Florence, Italy. Association for Computational Linguistics.

Kateřina Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. 2019. A test suite and manual evaluation of document-level NMT at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy. Association for Computational Linguistics.

Karin Sim Smith and Lucia Specia. 2018. Assessing crosslingual discourse relations in machine translation. *ArXiv*, abs/1810.03148.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Tereza Vojtěchová, Michal Novák, Miloš Klouček, and Ondřej Bojar. 2019. SAO WMT19 test suite: Machine translation of audit reports. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 481–493, Florence, Italy. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of EMNLP*.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. *ArXiv*, abs/2002.06823.