Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3704–3709
Marseille, 11–16 May 2020

A Test Set for Discourse Translation from Japanese to English

Masaaki Nagata, Makoto Morishita
NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai Seika-cho Souraku-gun Kyoto-fu 619-0237 Japan
{masaaki.nagata.et, makoto.morishita.gr}@hco.ntt.co.jp

Abstract

We made a test set for Japanese-to-English discourse translation to evaluate the power of context-aware machine translation. For each discourse phenomenon, we systematically collected examples where the translation of the second sentence depends on the first sentence. Compared with a previous study on test sets for English-to-French discourse translation (Bawden et al., 2018), we needed different approaches to make the data because Japanese has zero pronouns and represents different senses in different characters. We improved the translation accuracy using context-aware neural machine translation, and the improvement mainly reflects the betterment of the translation of zero pronouns.

Keywords: discourse, translation, test

1. Introduction

Translation accuracy for sentences has been greatly improved by neural machine translation (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017). Although the accuracy of sentence-level machine translation is approaching human parity (Hassan et al., 2018), human evaluators prefer the translations of human translators over those of machine translation in document-level evaluations (Samuel Läubli, 2018). For further improvement, research using discourse information is gaining attention (Tiedemann and Scherrer, 2017; Bawden et al., 2018; Voita et al., 2018; Maruf and Haffari, 2018; Miculicich et al., 2018; Voita et al., 2019a).

The easiest method to use discourse information is to concatenate both the previous and current sentences with a special token <CONCAT> as a separator and translate them using an ordinary sentence-based translation system (Tiedemann and Scherrer, 2017). This method, which is called 2-to-2, is known to be a reasonably strong baseline for discourse translation (Bawden et al., 2018; Voita et al., 2018). In this context, ordinary sentence-based translation is called 1-to-1. Some recent proposals have focused on multiple-encoders, where the input sentence and the context have different encoders. Bawden et al. (2018) used an RNN-based encoder-decoder and Voita et al. (2018) used the Transformer (Vaswani et al., 2017) for multiple-encoders. Others used a cache-based model to exploit document-level information (Wang et al., 2017; Kuang et al., 2018; Tu et al., 2018).

One problem in the study of discourse translation is that by only using automatic accuracy measures such as BLEU, we cannot distinguish between what is solved and not solved by the proposed method. Bawden et al. (2018) therefore proposed discourse test sets for English-to-French translation to evaluate the power of generating appropriate sentences based on the context information using the framework of a contrastive test set (Sennrich, 2017). For discourse phenomena, they targeted coreference and coherence, and manually created a test set based on real examples in bilingual corpora. Voita et al. (2019b) made a contrastive test set for English-to-Russian discourse translation.

To make Japanese-to-English discourse translation test sets, we also targeted on coreference and coherence. We first looked for examples in bilingual corpora. However, since we found that this approach is inefficient, we chose a different approach using linguistically annotated corpora.

The following are this paper's contributions:

- we created a novel test set for Japanese-to-English discourse translation[1].
- we proposed a novel approach to create discourse translation tests: tests for coreferences were made from monolingual annotated databases and those for coherences were made from bilingual dictionaries.
- using the proposed test set, we show that improvement of the context-aware neural machine translation is mainly brought by the betterment of the translation of Japanese zero pronouns.

In the following sections, we first describe our Japanese-to-English discourse translation test set. We then describe our experiment results on discourse translation and analyze them using our proposed discourse test set.

2. Contrastive Test Set for Discourse Translation

To make a contrastive test set for Japanese-to-English discourse translation, we first made a pair of Japanese sentences and their English translations, where the English translation of the second Japanese sentence depends on the first Japanese sentence. We then made an English sentence by adding to the second English sentence a minor error caused by ignoring the context provided by the first sentence. We tried to set the correct answer rate of the contrastive test to 50%, which means that the translation model estimates the correct and incorrect choices as equally plausible if it does not know the context. The target linguistic phenomena are coreference and coherence and the target number of tests was about 1000.

[1] Our discourse test set is available at `https://github.com/nttcslab-nlp/discourse-mt-test-sets/`.

Source:	
Context:	田中さんはセンター試験の成績が良かった。
Input:	多分(*pro* が)東大を受験するだろう。
Target:	
Context:	Mr. Tanaka got good results in the center exam.
Correct:	Maybe he will take the entrance exam for the University of Tokyo.
Incorrect:	Maybe I will take the entrance exam for the University of Tokyo.

(a) Zero pronoun

Source:	
Context:	ソファのそばには木製の椅子がある。
Input:	レンズとピンセットが椅子に乗っている。
Target:	
Context:	Beside the couch was a wooden chair.
Correct:	A lens and a forceps was lying upon the seat.
Incorrect:	A lens and a forceps was lying upon a seat.

(b) Article

Figure 1: Discourse tests for coreference

2.1. Coreference

2.1.1. Zero Pronouns

For the coreferences, we focus on the target linguistic elements without counterparts in the source language (Japanese zero pronouns and English articles) because the generation of such linguistic elements is dependent on the context. In Japanese, when subjects and objects are understood from the context, they are usually omitted. These omitted elements are called *zero pronouns* because they resemble pronouns without surface forms. To translate a Japanese zero pronoun, we have to determine the person, the number, and the grammatical function of an English pronoun.

In the example in Figure 1(a), the default interpretation of the omitted subject for the action verb is the first person, but the context provided by the previous sentence changes it to the third person.

Bawden et al. (2018) created discourse test sets inspired by real examples found in OpenSubtitles. Although we also tried the same procedure, we found it inefficient because correctly analyzing Japanese zero pronouns is difficult, even for native Japanese speakers. We therefore adopted the following procedure. First, we selected a Japanese sentence with zero pronouns from a corpus with annotation on zero pronouns, such as Keyaki Treebank [2] (Butler et al., 2012). We then translated the sentence by Google Translate to obtain the most likely English translation for the Japanese zero pronoun because their En-Ja dataset is larger than ours[3]. We then selected an English pronoun for an incorrect English translation and made two preceding Japanese sentences, one for the correct English pronoun and another for the incorrect one.

For coreference, Bawden et al. (2018) defined "semi-correct" to denote when second sentence's translation is correct in terms of the target side context when the first sentence's translation is incorrect. They made a set of four sentences for a test: correct/incorrect, semi-correct/incorrect. We only made two sentences (correct/incorrect) because the consistency on the target side context can be evaluated by coherence tests.

2.1.2. Articles

Figure 1(b) shows an example of a test for an article. Since the same entity is mentioned in the first sentence, the article in the second sentence must be the definite article "the". Note that the surface forms of 椅子 on the Japanese side are identical, while those on the English side are different: chair and seat[4].

We made the following tests for articles (definite/indefinite) with two types of previous sentences: with antecedents (including bridge reference) and without. We tried to make their proportion 50:50. As in the case for coreference, since looking for examples from the translation corpus is inefficient, we used annotated corpora for grammatical error correction (GEC) such as the Cambridge Learner Corpus First Certificate (CLC FCE) dataset[5] and coreference resolution such as OntoNotes 5.0 (Hovy et al., 2006). In particular, we found GEC examples useful when "the" was changed to "a" for making a natural pair of sentences without a definite reference:

> In our country, there are rules that everyone has to follow, and recently a new rule was added. We aren't allowed to use a (*the) mobile phone in class.

Finding definite references and their antecedents in the coreference resolution data is easy. However, as we described in the experiment section, it is difficult to make the correct answer rate be 50%, because most articles can be predicted using language models.

2.2. Coherence and Cohesion

For coherence and cohesion, we followed the previously defined classification of disambiguation and alignment (and repetition) (Bawden et al., 2018) because they are language independent. Figure 2(a) shows an example of a test for disambiguation, where "すごい人" can be translated into either "a lot of people" or "a great man."

In general, tests for disambiguation must satisfy the following three conditions:

- the source sentence has an ambiguous word;
- its senses are translated into different target words;
- the choice of target words depends on either the source or the previous target sentence.

Since Japanese uses Kanji (ideograms imported from Chinese), relatively few Japanese words have multiple senses

[2] http://www.compling.jp/keyaki/

[3] "These datasets are two to three orders of magnitude larger than the WMT datasets." (Johnson et al., 2017)

[4] This is an example of a bridge reference.

[5] https://www.ilexir.co.uk/datasets/index.html

Source:
Context: 昨日、渋谷へ行った。
Input: すごい人だった。
Target:
Context: I went to Shibuya yesterday.
Correct: There are a lot of people.
Incorrect: He is a great man.

(a) Disambiguation (Shibuya, Tokyo is famous for the world's most crowded intersection!)

Source:
Context: いい時計ですね。
Input: この時計は父の形見なんです。
Target:
Context: It's a nice clock.
Correct: This clock is a memento of my father.
Incorrect: This watch is a memento of my father.

(b) Alignment

Figure 2: Discourse tests for coherence and cohesion

and different English translations, because different senses are represented by different characters. We therefore used multiple Japanese-to-English dictionaries to collect ambiguous Japanese words and created tests from scratch. Figure 2(b) shows an example of an alignment test, where "時計" can be translated as either "clock" or "watch."

The conditions for alignment tests are slightly different from those for disambiguation:

- the source sentence has an ambiguous word;
- its senses are translated into different target words that are not interchangeable;
- the previous source sentence has the same ambiguous word and can be translated into the same set of target words depending on the ambiguous source word's sense.

Since the target translations of the ambiguous source word are not interchangeable, the target translation must be identical in the first and second sentences. Note that some grounding is required to decide the correct answer. For example in Figure 2(b), whether it is a clock or a watch depends on the object in the real word referred to by the linguistic expression "時計".

3. Experiments

3.1. Datasets

Table 1 shows the eight Japanese-English datasets used for our discourse translation experiments. They include spoken language datasets of about 2M sentences and written language datasets of about 1M sentences.

IWSLT-2017 (Cettolo et al., 2017) is a dataset for Japanese-English Tasks of the International Workshop on Spoken Language Translation and consists of the transcriptions of TED Talks and their translations.

OpenSubtitles2018 (Lison et al., 2018) is a collection of movie subtitles and their translations. Global Voices is a multilingual corpus created from Global Voices websites that translate social media and blogs (Prokopidis et al., 2016). Both are available from OPUS (Tiedemann, 2012).

Dataset	Split	sents	len(ja)	len(en)
IWSLT2017	train	218,174	22.3	20.6
(TED Talks)	dev	2,577	21.8	19.1
	test	2,357	22.5	19.5
OpenSubtitles2018	train	2,077,430	7.6	8.5
(movie subtitles)	dev	3,245	9.0	7.7
	test	2,901	6.9	8.9
GlobalVoices	train	29,508	27.8	21.6
(blog)	dev	9,426	28.4	22.0
	test	8,148	28.1	21.8
HiraganaTimes	train	16,472	24.6	22.0
Books	dev	2,792	22.4	20.3
(book)	test	2,537	23.4	21.3
HiraganaTimes	train	189,925	24.7	21.1
(magazine)	dev	5,385	21.3	19.8
	test	5,004	21.1	19.9
NICT_align	train	103,417	20.5	14.9
(book)	dev	4,279	21.0	15.1
	test	3,212	17.8	13.1
Wikipedia_Kyoto	train	480,778	23.4	24.9
(Wikipedia)	dev	1,257	20.2	19.9
	test	1,287	21.3	21.7
Yomiuri_editorial	train	283,710	27.2	28.2
(newspaper	dev	3,002	23.3	24.8
editorials)	test	3,014	24.3	26.4
All	train	3,399,414	14.0	14.3
	dev	31,963	22.4	19.1
	test	28,460	22.0	19.4

Table 1: Datasets statistics

Hiragana Times Books are a collection of bilingual books in Japan and the Hiragana Times is a monthly bilingual magazine for Japanese learners.[6] Both can be purchased at the publishing company.

The English-Japanese Translation Alignment Data (Utiyama and Takahashi,) (NICT_align) is a collection of publicly available books from such archives as Aozora Bunko[7] and Project Gutenberg[8]. Their document and sentence alignments were provided by NICT.

The Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles[9] (Wikipedia_Kyoto) is a collection of Japanese Wikipedia articles on Kyoto and their translations into English by NICT. This is the source of the Kyoto Free Translation Task (KFTT)[10], one of the most popular Japanese-English translation benchmarks. We used the same train, dev, and test split of documents for creating the discourse translation dataset.

Yomiuri_editorial is a collection of newspaper editorials from the Yomiuri Shimbun and their translations published

[6] https://www.hiraganatimes.com/

[7] https://www.aozora.gr.jp/

[8] https://www.gutenberg.org/

[9] https://alaginrc.nict.go.jp/WikiCorpus/index_E.html

[10] http://www.phontron.com/kftt/index.html

	RNN		Transformer	
Dataset	1-to-1	2-to-2	1-to-1	2-to-2
IWSLT2017	12.19	**12.26**	**17.72**	17.67
OpenSubtitles2018	12.23	**12.67**	15.07	**15.48**
GlobalVoices	10.66	**10.80**	14.86	**14.96**
HiraganaTimes_books	12.91	**13.23**	18.03	**18.56***
HiraganaTimes	13.23	**13.32**	19.22	**19.69***
NICT_align	9.36	**9.64**	12.92	**13.38**
Wikipedia_Kyoto	23.09	**23.12**	27.80	**28.08**
Yomiuri_Editorial	14.53	**15.26***	**21.89**	21.76
All	12.77	**13.02***	17.87	**18.07***

Table 2: Translation accuracies of each dataset for 1-to-1 and 2-to-2 models by two translation methods. * indicates statistically significant difference ($p \leq 0.01$).

at The Japan News (formerly the Daily Yomiuri), which is the newspaper's English edition. We purchased CD-ROMs published for research purposes [11], and obtained document and sentence alignments using the algorithm of Utiyama and Isahara (2003).

3.2. Tools

For preprocessing, the English sentences are tokenized and lowercased by the scripts in Moses toolkit (Koehn et al., 2007). Japanese sentences were normalized by NFKC (a unicode normalization form) and word segmented by MeCab[12] with UniDic. Both Japanese and English sentences were further tokenized into subwords using byte pair encoding (Sennrich et al., 2016) with 32k shared merge operations.

For neural machine translation, we used OpenNMT-lua[13] for RNN encoder-decoder (Luong et al., 2015) and fairseq toolkit[14] for the Transformer (Vaswani et al., 2017). We used default settings unless otherwise specified. The translation accuracy was measured by BLEU (Papineni et al., 2002) using multi-bleu.perl in the Moses toolkit.

3.3. Context Boundaries

Unlike spoken language datasets, written language datasets have obvious structures including documents, sections, and paragraphs, all of which can be candidates for context boundaries. To the best of our knowledge, scant research has addressed context boundaries in previous works on discourse translation, probably because they are based on spoken language datasets such as OpenSubtitles.

As a preliminary experiment, we compared the translation accuracies of two different context boundaries (document (file) and paragraph) with a 2-to-2 discourse translation model and found virtually no differences in translation accuracies. We therefore used document boundaries for the contexts in the following experiments.

3.4. Translation Accuracies

We made 1-to-1 and 2-to-2 translation models from all datasets. Table 2 shows the translation accuracies (BLEUs) of the different test sets and two translation methods. In general, the 2-to-2 models outperformed the 1-to-1 models with a small but statistically significant margin. Transformer's accuracy was significantly higher than attention-based RNN encoder-decoder.

3.5. Discourse Test Set Scores

We obtained the translation scores (log probability) for the correct and incorrect sentences of the discourse test sets by forced decoding and calculated the proportion of the tests where the scores of the correct sentences exceeded those of the incorrect sentences. Table 3 shows the correct answer rate of each test category for the 1-to-1 and 2-to-2 models of the two translation methods.

As for the Transformer, the 2-to-2 models consistently outperformed the 1-to-1 models in all the test categories. In particular, the correct answer rate of pronoun is greatly improved compared with other categories. As for the RNN encoder-decoder, pronoun is the only category with significant improvement. The correct answer rate of the 1-to-1 model for articles was significantly higher than the designed baseline (50%). We assume this is because there are many cases where correct articles can be predicted based on the local context in the current sentence regardless of the previous sentence.

For pro-drop languages like Japanese and Chinese, zero pronoun was known to be one of the most difficult problems and many specific extensions for baseline translation methods have been discussed in previous research (Taira et al., 2012; Kudo et al., 2014; Takeno et al., 2016; Wang et al., 2016; Wang et al., 2018). However, it seems that context-aware neural machine translation can handle Japanese zero pronouns just as effectively as overt pronouns in English-to-Russian translation (Voita et al., 2018). To the contrary, there is room to improve the handling of coherence in simple context-aware models, such as the 2-to-2 model.

3.6. Sample Usage of Test Set

Figure 3 shows an example of a test for pronoun translation. In "Input:", the previous and current sentences are concatenated by a special token, <CONCAT>. The current sentence has two zero pronouns: subjects of the subordinate clause and the matrix sentence. A zero pronoun is indicated by *pro*, and the subject is indicated by が, which is the Japanese subject case marker. Comparing "Correct:" with "Incorrect:", the subject of the subordinate clause was changed from "she" to "you." Since the log probability (-27.22) for the correct sentence is larger than that for the incorrect sentence (-29.67), the example is categorized as correct. "Translation:" is the output of the 2-to-2 model, where both Japanese zero pronouns are correctly translated into English (overt) pronouns.

[11] http://www.nichigai.co.jp/sales/corpus.html

[12] http://taku910.github.io/mecab/

[13] https://opennmt.net

[14] https://github.com/pytorch/fairseq

Test Set		tests	RNN 1-to-1	RNN 2-to-2	Transformer 1-to-1	Transformer 2-to-2
Coreference	article	330	**0.76**	0.75	0.78	**0.82**
	pronoun	220	0.53	**0.70**	0.56	**0.82**
Coherence	disambiguation	378	0.50	**0.52**	0.50	**0.56**
	alignment	73	0.49	**0.50**	0.50	**0.59**

Table 3: Correct answer rate of each linguistic test category for 1-to-1 and 2-to-2 models

2-to-2	
Input:	あの彼女、今度うちの学校に来るらしいよ。<CONCAT> けど、(*pro* が)$_1$ いつ来るか、(*pro* が)$_2$ わからない。
Correct:	that girl seems to come to our school soon . <CONCAT> but i$_2$ don 't know when she$_1$ comes . -27.22
Incorrect:	that girl seems to come to our school soon . <CONCAT> but i$_2$ don 't know when you$_1$ come . -29.67
Translation:	she 's coming to my school next time . <CONCAT> but i$_2$ don 't know when she$_1$'s coming . -8.25
1-to-1	
Input:	けど、(*pro* が)$_1$ いつ来るか、(*pro* が)$_2$ わからない。
Correct:	but i$_2$ don 't know when she$_1$ comes . -8.48
Incorrect:	but i$_2$ don 't know when you$_1$ come . -7.44
Translation:	but i$_2$ don 't know when . -3.24

Figure 3: Examples for input source, correct target, incorrect target, and output target sentences for 1-to-1 and 2-to-2 models with their translation scores

For the 1-to-1 model, the log probability (-8.48) of the correct sentence is smaller than the incorrect sentence (-7.44), which is categorized as a wrong answer. The translation output abandons the translation of the subordinate clause, which is a typical behavior of neural machine translation.

4. Conclusions

We developed a contrastive test set for evaluating the power of discourse translation models for Japanese-to-English translation and found that the 2-to-2 discourse translation model's improvement is mainly caused by better translation of Japanese zero pronouns. As was also shown in Kimura et al. (2019), Japanese zero pronouns can basically be effectively handled by context-aware neural machine translation. In future work, we want to build a test set for English-to-Japanese discourse translation to focus on Japanese empathy and honorifics. We also want to develop a more sophisticated context-aware neural machine translation method that can appropriately handle coherence.

Moreover, we want to expand our test set to a similar size of a previous work (Müller et al., 2018). They built a large-scale test set from German-English bilingual texts using coreference resolution and word alignment tools. However, to build a large-scale test set for Japanese zero pronouns, we have to develop accurate tools for Japanese empty category detection (zero pronoun identification) and Japanese coreference resolution, which remains open problems.

5. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the ICLR-2015*.

Bawden, R., Sennrich, R., , Birch, A., and Haddow, B. (2018). Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the NAACL-2018*, pages 1304–1313.

Butler, A., Hotta, T., Otomo, R., Yoshimoto, K., Zhou, Z., and Zhu, H. (2012). Keyaki treebank: Phrase structure with functional information for japanese. In *Proceedings of Text Annotation Workshop*.

Cettolo, M., Federico, M., Bentivogli, L., Niehues, J., Stüker, S., Sudoh, K., Yoshino, K., and Federmann, C. (2017). Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the IWSLT-2017*, pages 2–14.

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 *Proceedings of the EMNLP-2014*, pages 1724–1734.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic chinese to english news translation. arXiv:1803.05567.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: The 90% solution. In *Proceedings of the NAACL-2006*, pages 57–60.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, pages 339–351.

Kimura, R., Iida, S., Cui, H., Hung, P.-H., Utsuro, T., and Nagata, M. (2019). Selecting informative context sentence by forced back-translation. In *Proceedings of the MT Summit XVII*, page 162–171.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL-2007*, pages 177–180.

Kuang, S., Xiong, D., Luo, W., and Zhou, G. (2018). Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the COLING-2018*, page 596–606.

Kudo, T., Ichikawa, H., and Kazawa, H. (2014). A joint inference of deep case analysis and zero subject generation for japanese-to-english statistical machine translation. In *Proceedings of the ACL-2014*, pages 557–562.

Lison, P., Tiedemann, J., and Kouylekov, M. (2018). Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the LREC-2018*, pages 1742–1748.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the EMNLP-2015*, pages 1412–1421.

Maruf, S. and Haffari, G. (2018). Document context neural machine translation with memory networks. In *Proceedings of the ACL-2018*, page 1275–1284.

Miculicich, L., Ram, D., Pappas, N., and Henderson, J. (2018). Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the EMNLP-2018*, pages 2947–2954.

Müller, M., Rios, A., Voita, E., and Sennrich, R. (2018). A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of WMT-2018*, pages 61–72.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the ACL-2002*, pages 311–318.

Prokopidis, P., Papavassiliou, V., and Piperidis, S. (2016). Parallel global voices: a collection of multilingual corpora with citizen media stories. In *Proceedings of the LREC-2016*, pages 900–905.

Samuel Läubli, Rico Sennrich, M. V. (2018). Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the EMNLP-2018*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the ACL-2016*, pages 1715–1725.

Sennrich, R. (2017). How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. In *Proceedings of the EACL-2017*, pages 376–382.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Proceedings of the NIPS-2014*, pages 3104–3112.

Taira, H., Sudoh, K., and Nagata, M. (2012). Zero pronoun resolution can improve the quality of j-e translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-2012)*, pages 111–118.

Takeno, S., Nagata, M., and Yamamoto, K. (2016). Integrating empty category detection into preordering machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT-2016)*, pages 157–165.

Tiedemann, J. and Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the LREC-2012*, pages 2212–2218.

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.

Utiyama, M. and Isahara, H. (2003). Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the ACL-2003*, pages 72–79.

Utiyama, M. and Takahashi, M.). English-japanese translation alignment data.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the NIPS 2017*, pages 5998–6008.

Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (2018). Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the ACL-2018*, pages 1264–1274.

Voita, E., Sennrich, R., and Titov, I. (2019a). Context-aware monolingual repair for neural machine translation. In *Proceedings of the EMNLP-IJCNLP-2019*, page 877–886.

Voita, E., Sennrich, R., and Titov, I. (2019b). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the ACL-2019*, page 1198–1212.

Wang, L., Tu, Z., Zhang, X., Li, H., Way, A., and Liu, Q. (2016). A novel approach to dropped pronoun translation. In *Proceedings of the NAACL-HLT-2016*, pages 983–993.

Wang, L., Tu, Z., Way, A., and Liu, Q. (2017). Exploiting cross-sentence context for neural machine translation. In *Proceedings of the EMNLP-2017*, page 2826–2831.

Wang, L., Tu, Z., Way, A., and Liu, Q. (2018). Learning to jointly translate and predict dropped pronouns with a shared reconstruction mechanism. In *Proceedings of the EMNLP-2018*, pages 2997–3002.