Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3630–3634
Marseille, 11–16 May 2020

Evaluation Dataset for Zero Pronoun in Japanese to English Translation

Sho Shimazu, Sho Takase, Toshiaki Nakazawa, Naoaki Okazaki
Tokyo Institute of Technology, The University of Tokyo
{sho.shimadu, sho.takase}@nlp.c.titech.ac.jp, nakazawa@logos.t.u-tokyo.ac.jp, okazaki@c.titech.ac.jp

Abstract

In natural language, we often omit some words that are easily understandable from the context. In particular, pronouns of subject, object, and possessive cases are often omitted in Japanese; these are known as zero pronouns. In translation from Japanese to other languages, we need to find a correct antecedent for each zero pronoun to generate a correct and coherent translation. However, it is difficult for conventional automatic evaluation metrics (e.g., BLEU) to focus on the success of zero pronoun resolution. Therefore, we present a hand-crafted dataset to evaluate whether translation models can resolve the zero pronoun problems in Japanese to English translations. We manually and statistically validate that our dataset can effectively evaluate the correctness of the antecedents selected in translations. Through the translation experiments using our dataset, we reveal shortcomings of an existing context-aware neural machine translation model.

Keywords: machine translation, zero pronoun, language resources

1. Introduction

Neural encoder-decoder models have achieved high BLEU scores on the machine translation task (Bahdanau et al., 2015; Vaswani et al., 2017). However, typical machine translation models translate each sentence without surrounding sentences into consideration since they assume a one-to-one correspondence between a source sentence and a target sentence. Moreover, the popular metrics evaluate the quality of each translated sentence regardless of its context (Papineni et al., 2002). These configurations cannot deal with certain linguistic phenomena that depend on the surrounding sentences, such as co-reference (Le Nagard and Koehn, 2010; Guillou, 2012) and lexical cohesion (Bawden et al., 2018).

In contrast, several studies extended a widely used encoder-decoder model to consider multiple sentences (Tiedemann and Scherrer, 2017; Voita et al., 2018; Voita et al., 2019). Although the methods improved the translation quality, it is difficult to demonstrate whether these approaches correctly addressed the context-dependent linguistic phenomena because few test data focus on the phenomena.

Voita et al. (2018) evaluated co-reference chains of translations employing automatic co-reference resolution (Manning et al., 2014) and human annotation in English-Russian. Bawden et al. (2018) presented manually constructed test sets to evaluate co-reference and cohesion in English-French translations. Voita et al. (2019) analyzed outputs of a recent neural machine translation model on English-Russian to investigate an error depending on context sentences. They also presented test sets for three phenomena including deixis, lexical cohesion, and ambiguity which depends on an ellipsis to avoid repeating the same phrase. These studies contributed to evaluating context-dependent translation, but they did not consider the zero pronoun problem.

Zero pronoun, which is an omitted pronoun, is an important issue in machine translation because a model needs to complement such ellipsis during translation. Some European languages such as Portuguese and Spanish allow an ellipsis of subject but it is easy to complement the omitted pronoun without context sentences on such languages because a word depending on the omitted pronoun includes a marker such as the case. In contrast, Japanese is regarded as one of the most difficult language to translate in terms of zero pronoun phenomenon because subject, object, and possessive cases are often omitted in Japanese and Japanese words usually have no inflectional forms that depend on the omitted pronoun (Taira et al., 2012; Kudo et al., 2014).

Figure 1 shows an example. This figure indicates two Japanese sentences (JP) and their translations into Spanish (ES) and English (EN). Since some words are omitted in Japanese (i.e., “kare wo”, “kare wa”, and “kare no”), we have to complement them (“him”, “he”, and “his” are English words corresponding to the omitted words in the Japanese side) during translation. However, even state-of-the-art methods for Japanese zero pronoun resolution achieved an F1 score of only approximately 60% (Matsubayashi and Inui, 2018; Shibata and Kurohashi, 2018; Kurita et al., 2018). On the other hand, Spanish sentences also contain an ellipsis (“él” corresponding to “him” and “he”) but we can complement it easily based on the related word “su” which is corresponding to “his”. Thus, machine translation considering zero pronoun is a challenging task and zero pronoun in Japanese is a difficult issue in translation from Japanese to other languages.

In this paper, we construct a test set to evaluate zero pronoun resolution in Japanese-English translations. The test set contains a pair of Japanese-English sentences and Japanese context sentences. A machine translation model is expected to find a correct antecedent for a zero pronoun from context sentences. We also evaluate previous context-aware translation models on our constructed test set to explore future direction.

2. Evaluation of Zero Pronoun

2.1. Background

Bawden et al. (2018) constructed test sets to evaluate coherence and cohesion in contextualized translation. Their test sets consist of four components: a sentence as a source of translation (current sentence), a previous sentence to the current sentence, which is used as context for translation, and the correct and incorrect translations of the current sen-

JP: 母も父も助けも来ず、通勤途中の救急医療隊員に見つけられるまで**彼**は1人ぼっちでした。// 我々が (**彼を**)発見した時、(**彼は**) (**彼の**)自転車にまたがったままでした。

haha mo chichi mo tasuke mo ko zu, tsukin tochu no kyukyuiryo taiin ni mitsuke rareru made **kare** ha 1 ri bocchi deshi ta. // wareware ga (**kare wo**) hakken shi ta toki, (**kare wa**) (**kare no**) jitensha ni matagaltu ta mama deshi ta.

ES: ***El*** *estaba completamente solo, sin su madre ni su padre y sin ayuda hasta que finalmente fue visto por una paramédica que caminaba a* ***su*** *trabajo. // Cuando nosotros lo encontramos a (* ***él*** *) todavía estaba en* ***su*** *bicicleta.*

EN: ***He*** *was all alone no mom no dad no help until finally* ***he*** *was spotted by a paramedic on her way to work. // When we found* ***him***, ***he*** *was still on* ***his*** *bike.*

Figure 1: Example of the zero pronoun problem. These sentences represent a conversation between a reporter and a paramedic.

tence. The current sentence contains a word for which we need the context sentence to translate it correctly, such as ambiguous anaphoric pronouns *la* and *le* in French. Thus, a translation model needs to pay attention to the context sentence to output the correct translation. They evaluated whether each model selects the correct translation based on a score computed by the model.

Following Bawden et al. (2018), we construct a test set that consists of four components: a current sentence, context, a correct translation, and an incorrect translation. In our test set, since the current sentence contains zero pronoun, translation models need to detect the correct antecedent from the context sentences.

2.2. Test Set Construction

To avoid contaminating the test set with noisy instances, we construct our test set with fully hand-crafted. Thus, the construction of our test set consisted of three steps: translation, zero pronoun detection, and incorrect instance construction.

To easily detect zero pronouns whose antecedent appears in previous sentences, we translated an existing corpus that contains manually annotated co-reference chain into another language. In addition, Taira et al. (2012) indicated that spoken language tends to contain more zero pronoun than literary language. Thus, we used the CNN broadcast conversation section of OntoNotes (Ralph et al., 2013; Pradhan et al., 2007) as a corpus that satisfies the above requirements. The corpus consists of 5,341 English sentences. We asked translation experts to avoid translating the sentence word for word and to instead translate each sentence in fluent Japanese to obtain zero pronoun instances.

The purpose of the zero pronoun detection step is to obtain Japanese sentences that contain a zero pronoun whose antecedent appears in previous sentences. We collected English sentences including a pronoun whose antecedent appears in previous sentences based on the co-reference chain. Then, we extracted the zero pronoun from the translations of the above English sentences.

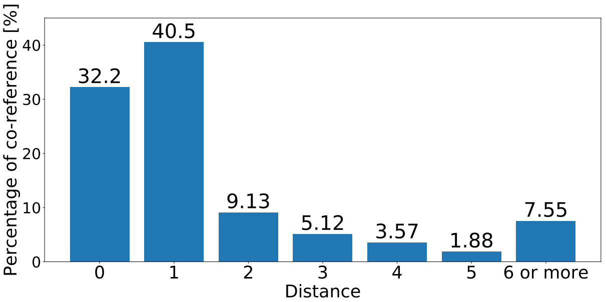


Figure 2: Statistics of distance from a current sentence to a previous sentence containing the antecedent of a pronoun in the current sentence. We regard the nearest sentence that contains the antecedent from the current sentence as the previous sentence.

Figure 2 shows statistics on distance between a current sentence and a previous sentence containing the antecedent of a pronoun in the current sentence[1]. This figure indicates that three previous sentences cover 80.7% of all co-referred antecedents. Thus, we used previous three sentences as the context in contrast to Bawden et al. (2018) which used only the previous one sentence as a context. In addition, we include the distance to the context sentence where the antecedent of a zero pronoun occurs. We represent this distance as d.

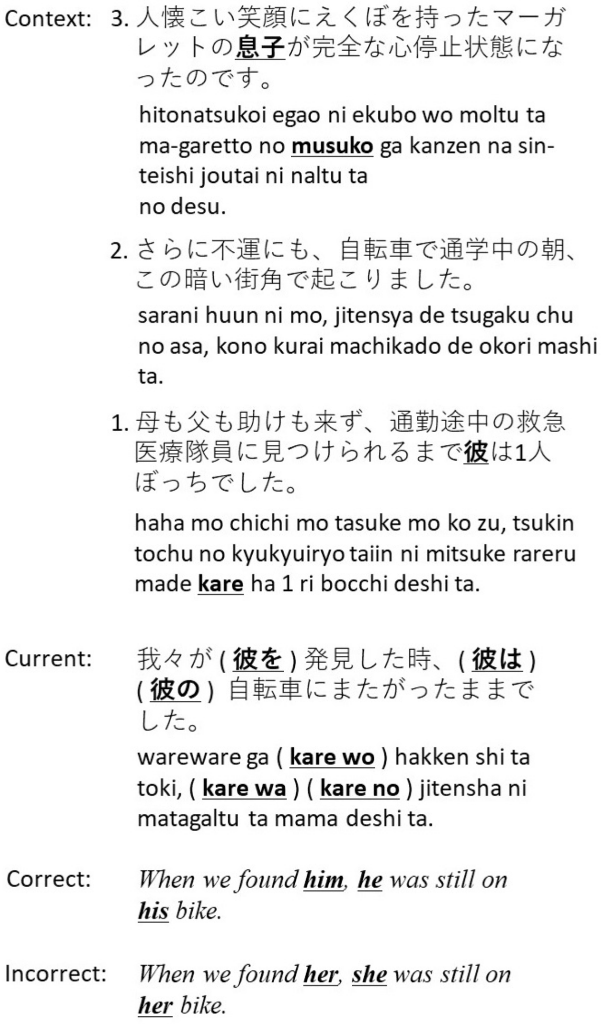
In the incorrect instance construction step, we created a contrastive instance that is an English sentence containing an incorrect pronoun. Thus, in the correct sentence (original English sentence), we replaced the pronoun corresponding to the zero pronoun in the translated Japanese with another pronoun randomly. Moreover, we replaced all pronouns tied to the co-reference chain in English. We also modified a subject-verb agreement if necessary. Figure 3 shows an example of our test set.

In addition to such contextualized translation, we also constructed $d = 0$ to evaluate whether a translation model handles intra-sentential pronouns coherently. Through the above procedure, we created 724 examples: 218 for $d = 0$, 362 for $d = 1$, 96 for $d = 2$, and 48 for $d = 3$.

After constructing the test set, we investigated the accuracy of humans with and without contexts to demonstrate that humans can select the correct sentence only if they have the context of the current sentence. We asked two people to select one sentence from two candidates. Moreover, in cases where they did not have enough context to judge, we asked them to answer "undecidable" and we regarded such instances as incorrect choice.

The bottom row of Table 1 shows the averaged accuracy of them. This table shows that human judgments achieved nearly 100% for $d = 0$ even if they had no contexts because context is not necessary in this case. However, for $d \geq 1$, human judgments without context sentences (HUMAN W/O CONTEXT) were much lower than random choices because

[1] We analyzed only co-reference cases because it is difficult to detect a zero pronoun instances in translated Japanese based on co-reference chain in English.



Context: 3. 人懐こい笑顔にえくぼを持ったマーガレットの**息子**が完全な心停止状態になったのです。

hitonatsukoi egao ni ekubo wo moltu ta ma-garetto no **musuko** ga kanzen na sin-teishi joutai ni naltu ta
no desu.

2. さらに不運にも、自転車で通学中の朝、この暗い街角で起こりました。

sarani huun ni mo, jitensya de tsugaku chu no asa, kono kurai machikado de okori mashi ta.

1. 母も父も助けも来ず、通勤途中の救急医療隊員に見つけられるまで**彼**は1人ぼっちでした。

haha mo chichi mo tasuke mo ko zu, tsukin tochu no kyukyuiryo taiin ni mitsuke rareru made **kare** ha 1 ri bocchi deshi ta.

Current: 我々が (**彼を**) 発見した時、(**彼は**) (**彼の**) 自転車にまたがったままでした。

wareware ga (**kare wo**) hakken shi ta toki, (**kare wa**) (**kare no**) jitensha ni matagaltu ta mama deshi ta.

Correct: *When we found **him**, **he** was still on **his** bike.*

Incorrect: *When we found **her**, **she** was still on **her** bike.*

Figure 3: Example of our test set. These sentences represent a conversation between a reporter and a paramedic. The context sentences are spoken by the reporter and the current sentence is spoken by the paramedic.

they answered "undecidable". Moreover, in these settings, kappa coefficients were also much lower than judgments with context sentences (HUMAN W/ CONTEXT). These results imply that humans cannot select consistent candidates if they have no contexts that are necessary for finding out the correct antecedents for zero pronouns.

In contrast, human judgments with context sentences achieved approximately 90% with high kappa coefficients in all settings. Therefore, we conclude that humans can detect the correct antecedents in our constructed test set only if they have sufficient context sentences. In other words, our test set is suitable for evaluating whether translation models resolve zero pronoun problem by using context sentences.

3. Experiments

In this section, we demonstrate the usefulness of the test set in various aspects. Firstly, we show that simple statistical baselines cannot choose the correct translation. Secondly, we report the performances of NMT models that can or cannot consider contextual information.

MAJORITY This method selects an English translation containing a more frequent pronoun. Since we use the frequency in training data, this method assesses the difficulty of resolving zero pronouns with only frequency information.

CO-OCCUR This method selects an English sentence based on co-occurrence statistics between a pronoun and surrounding words. We regarded the inner product of word embeddings trained by CBOW (Mikolov et al., 2013)[2] as the metric for co-occurrence because such embeddings can be interpreted as factorized results of a co-occurrence matrix (Levy and Goldberg, 2014). We define the co-occurrence score as follows:

$$\textsc{Co-occur}(\mathrm{x}, \mathrm{W}) = \sum_{w \in W \setminus x} v_x \cdot v_w, \tag{1}$$

where W is a set of words in the sentence, x is a target pronoun in the correct (or incorrect) sentence, and v_w is an embedding of a word w. We used two configurations for W: one was from the target side only (TARGET), i.e., W contains words in the correct (or incorrect) sentence and the other was from both of the target and the source (SOURCE+TARGET).

SINGLE-ENCODER We applied a widely-used LSTM encoder-decoder model (Luong et al., 2015) as the basic NMT model. We used the implementation of OpenNMT-py[3] with the same hyper-parameters as Bawden et al. (2018). In addition to translations from only the current sentence (1-TO-1), we evaluated the performance of the model with contexts. We concatenated context sentences with the current sentence to encode multiple sentences in this method.

MULTI-ENCODER We used the multi-encoder model with the hierarchical attention proposed by Bawden et al. (2018)[4] as a recent context-aware NMT model. This method comprises an additional encoder to incorporate context sentences. Since Bawden et al. (2018) used only one sentence as the context, they prepared only one additional encoder. In contrast, we prepared more encoders because we used at most three sentences as contexts. For this reason, we applied 3-TO-1 and 4-TO-1 in addition to 2-TO-1 in Bawden et al. (2018). We trained each method with the same hyper-parameters as Bawden et al. (2018).

3.1. Training Data

We have two requirements for the training corpus: the training corpus consists of conversation such as our constructed test set and contains context sentences for each Japanese-English sentence pair. In this paper, we used OpenSubtitles2016 (Lison and Tiedemann, 2016)[5] as a corpus that satisfies these requirements. We extracted Japanese-English sentence pairs from OpenSubtitles2016 with applying the

[2] We also applied Skip-gram but CBOW achieved better performance in our experiments.

[3] https://github.com/OpenNMT/OpenNMT-py

[4] They used GRU but we applied LSTM for comparison with the SINGLE-ENCODER.

[5] https://www.opensubtitles.org/ja

Model	$d = 0$	$d = 1$	$d = 2$	$d = 3$	Total
MAJORITY	41.7	47.8	50.0	50.0	46.4
CO-OCCUR (X, TARGET)	59.6	52.5	44.8	60.4	54.1
CO-OCCUR (X, SOURCE+TARGET)	67.0	56.1	57.3	75.0	60.8
SINGLE 1-TO-1	88.2±2.68	57.5±0.69	56.0±4.32	73.4±3.42	67.6±0.92
SINGLE 2-TO-1	89.4±1.91	68.0±1.50	57.5±4.14	**76.7± 3.06**	73.6±1.37
SINGLE 3-TO-1	89.6±2.24	68.8±1.27	**63.5± 2.87**	72.5±3.58	74.6±0.97
SINGLE 4-TO-1	90.0± 2.04	69.2±0.87	63.3±2.67	74.2±1.67	**75.0± 0.69**
MULTI 2-TO-1	89.8±0.84	67.8±0.90	56.7±1.06	75.8±2.12	73.5±0.29
MULTI 3-TO-1	89.0±2.05	**69.3± 1.55**	60.8±2.68	75.0±4.56	74.2±1.44
MULTI 4-TO-1	**90.0± 1.24**	69.2±1.01	62.3±2.75	76.7± 5.50	75.0± 1.14
HUMAN W/O CONTEXT	98.4*	27.3	26.0	38.5	49.3
HUMAN W/ CONTEXT	98.6*	93.5*	92.2*	86.5*	94.4*

Table 1: Accuracy of each method on our constructed test set. For NMT models, the averaged accuracy and standard deviation of five NMT models trained with different initial values are presented. For human evaluation, we add * to scores whose kappa coefficient was more than 0.7.

Model	BLEU
SINGLE 1-TO-1	17.0±0.54
SINGLE 2-TO-1	16.4±0.31
SINGLE 3-TO-1	15.9±0.14
SINGLE 4-TO-1	15.5±0.50
MULTI 2-TO-1	17.1±0.35
MULTI 3-TO-1	17.2±0.68
MULTI 4-TO-1	**17.5±0.39**

Table 2: BLEU score of each method on the test set extracted from OpenSubtitles2016.

pre-processing script presented by Bawden et al. (2018)[6]. Moreover, we split the extracted pairs to 150M, 5k, and 5k pairs for training, development, and test sets respectively while taking attention to remaining their contexts. We used the test set from OpenSubtitles2016 to calculate the BLEU score of each translation method because our constructed test set might be slightly small for the calculation. We used BPE (Sennrich et al., 2016) for vocabulary construction with 30k as merge operation.

3.2. Results

Table 1 shows the accuracy of each method on our constructed test set. This table implies that MAJORITY is almost equal to random choice since its accuracy is not higher than 50%. CO-OCCUR was superior to MAJORITY for all distances except for $d = 2$. This result indicates that a pronoun tends to co-occur with some tokens.

SINGLE 1-TO-1 achieved better accuracy than CO-OCCUR (X, SOURCE+TARGET) for $d = 0$ but does not on other distances. In contrast, the methods that encode contexts outperformed SINGLE 1-TO-1 for $d \geq 1$. These results imply that these methods can find out the correct antecedent from contexts during translation. On the other hand, there is no remarkable difference between SINGLE and MULTI in terms of the accuracy when each method use the same contexts.

Table 2 shows the BLEU scores of SINGLE and MULTI models on the test set extracted from OpenSubtitles2016. Table 2 indicates that the BLEU score of SINGLE dropped depending on the increase of context sentences. This tendency is also reported in Bawden et al. (2018).

In contrast, all MULTI models achieved better BLEU score than SINGLE 1-TO-1 and improved the score depending on the increase of the contexts. Therefore, we consider that MULTI is superior to SINGLE based on the results described in Table 1 and Table 2. Moreover, we should use as large number of context sentences as possible to achieve high BLEU score.

However, the accuracy of MULTI were still much lower than HUMAN W/ CONTEXT. To improve the accuracy of NMT models, Bawden et al. (2018) suggested predicting contexts. As an example of our future work, we consider that we address zero pronoun resolution in encoder side explicitly such as multi-task learning.

For $d = 3$, all models except for MAJORITY achieved high accuracy. In the correct sentences of this distance, the pronoun “I” is much more frequent than others. We consider that this unbalanced frequency causes such high accuracy. Since we might be able to fix this unbalance by increasing the size of the test set, we will explore cheaper method for augmentation with the identical quality to construction in this paper as a future work.

4. Conclusion

We presented the test set for the zero pronoun problem in Japanese-English translation. Through human evaluation, we show that our test set is suitable for evaluating whether a translation model detects correct antecedents from context sentences. In addition, we indicated that simple statistical methods cannot solve our test set. Moreover, experimental results indicated that there is a large gap between the accuracy of recent neural encoder-decoders and human score. Thus, our test set could potentially be used as an indicator to help in the development of context-aware NMT models.

Acknowledgements

The research results have been achieved by ”Research and Development of Deep Learning Technology for Advanced

[6] https://github.com/rbawden/PrepCorpus-OpenSubs

Multilingual Speech Translation", the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

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 International Conference on Learning Representations*, pages 1–15.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (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*, pages 1304–1313.

Guillou, L. (2012). Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10.

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 52nd Annual Meeting of the Association for Computational Linguistics*, pages 557–562.

Kurita, S., Kawahara, D., and Kurohashi, S. (2018). Neural adversarial training for semi-supervised Japanese predicate-argument structure analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 474–484.

Le Nagard, R. and Koehn, P. (2010). Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27*, pages 2177–2185.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th Language Resources and Evaluation Conference*, pages 1–16.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, pages 55–60.

Matsubayashi, Y. and Inui, K. (2018). Distance-free modeling of multi-predicate interactions in end-to-end Japanese predicate-argument structure analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 94–106.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007). OntoNotes: A Unified Relational Semantic Representation. In *International Conference on Semantic Computing*, pages 517–524.

Ralph, W., Martha, P., Mitchell, M., Eduard, H., Sameer, P., Lance, R., Nianwen, X., Ann, T., Jeff, K., Michelle, F., Mohammed, E.-B., Robert, B., and Ann, H. (2013). OntoNotes Release 5.0.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Shibata, T. and Kurohashi, S. (2018). Entity-centric joint modeling of Japanese coreference resolution and predicate argument structure analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 579–589.

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*, pages 111–118.

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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*, 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 56th Annual Meeting of the Association for Computational Linguistics*, pages 1264–1274.

Voita, E., Sennrich, R., and Titov, I. (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.