Contrastive Learning for Context-aware Neural Machine Translation Using Coreference Information

Yongkeun Hwang[1], Hyungu Yun[1], Kyomin Jung[1,2]
[1]Dept. of Electrical and Computer Engineering, Seoul National University, Seoul, Korea
[2]Automation and Systems Research Institute, Seoul National University, Seoul, Korea
{wangcho2k, youaredead, kjung}@snu.ac.kr

Abstract

Context-aware neural machine translation (NMT) incorporates contextual information of surrounding texts, that can improve the translation quality of document-level machine translation. Many existing works on context-aware NMT have focused on developing new model architectures for incorporating additional contexts and have shown some promising results. However, most existing works rely on cross-entropy loss, resulting in limited use of contextual information. In this paper, we propose CorefCL, a novel data augmentation and contrastive learning scheme based on coreference between the source and contextual sentences. By corrupting automatically detected coreference mentions in the contextual sentence, CorefCL can train the model to be sensitive to coreference inconsistency. We experimented with our method on common context-aware NMT models and two document-level translation tasks. In the experiments, our method consistently improved BLEU of compared models on English-German and English-Korean tasks. We also show that our method significantly improves coreference resolution in the English-German contrastive test suite.

1 Introduction

Neural machine translation (NMT) has achieved impressive performances on translation quality, due to the introduction of novel deep neural network (DNN) architectures such as encoder-decoder model (Cho et al., 2014; Sutskever et al., 2014), and self-attentional networks like Transformer (Vaswani et al., 2017). The state-of-the-art NMT systems are now even comparable with human translators in sentence-level performance.

However, there are a number of issues on document-level translation (Läubli et al., 2018). These include pronoun resolution across sentences (Guillou et al., 2018), which needs cross-sentential contexts. To incorporate such document-level contextual information, several methods for context-aware NMT have been recently proposed. Many of the works have focused on introducing new model architectures like multi-encoder models (Voita et al., 2018) for encompassing contextual texts of the source language. These works have shown significant improvement in addressing discourse phenomena such as anaphora resolution mentioned above, as well as moderate improvements in overall translation quality (Lopes et al., 2020).

Despite some promising results, most of the existing works have trained the model by minimizing cross-entropy loss, making the model rather exploit contextual information implicitly such as a form of regularization (Kim et al., 2019; Li et al., 2020). Data augmentation for context-aware NMT is also an important issue, despite that recent works have focused on back-translation (Huo et al., 2020).

In this paper, we propose a Coreference-based Contrastive Learning for context-aware NMT (CorefCL), a novel data augmentation and contrastive learning scheme leveraging coreference information. Cross-sentential coreference between the source and target sentence can be a good source of training signal for context-aware NMT since it occurs when one or more expressions refer to the same entity, thus reflects dependencies between the source and contextual sentences.

CorefCL starts by conducting automatic annotation of coreference between the source and contextual sentences. Then, the referred mentions on contextual sentences are corrupted by removing and/or replacing tokens to generate contrastive examples. With those contrastive examples, we introduce a contrastive learning scheme equipped with a max-margin loss which encourages the model to discriminate between the original examples and the contrastive ones. By doing so, CorefCL makes the model more sensitive to cross-sentential contextual information.

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 1135–1144
November 10–11, 2021.

We experimented with CorefCL on three English-German corpora and one English-Korean document-level corpus, including WMT, IWSLT TED talk, and OpenSubtitles'18 English-German subtitles translation task, and a web-crawled English-Korean subtitles translation. In all translation tasks, CorefCL consistently improves overall BLEU over baseline models without CorefCL. On experiments with three common context-aware model settings, we show that improvements by CorefCL are also model-agnostic. Finally, we show that the proposed method significantly improved the performance on ContraPro (Müller et al., 2018), an English-German contrastive coreference benchmark.

2 Related Works

2.1 Context-aware NMT

Context-aware machine translation has been vigorously studied to exploit the crucial context information in surrounding sentences. Recent works have shown that contextual information can help the model to generate not only more consistent but also more accurate translation (Smith, 2017; Voita et al., 2018; Müller et al., 2018; Kim et al., 2019).

In particular, Voita et al. (2018) introduced a context-aware Transformer model which is able to induce anaphora relations, Miculicich et al. (2018) showed that a model using cross-sentential contextual information significantly outperforms in document-level translation tasks, and Yun et al. (2020) insisted that context-aware models record the best performance especially in spoken language translation tasks where mandatory information tend to be sparse over multiple sentences.

The simplest method for context-aware machine translation is to concatenate all surrounding sentences and treat the concatenated sequence as a single sentence (Tiedemann and Scherrer, 2017). Although the concatenation strategy boosted Transformer architectures in multiple tasks (Tiedemann and Scherrer, 2017; Voita et al., 2018; Yun et al., 2020), it lagged behind efficiency as the Transformer architecture has limited long-range dependency (Tang et al., 2018).

To improve the efficiency, an additional encoder module is introduced to encode only the context sentences (Libovický and Helcl, 2017; Jean et al., 2017; Voita et al., 2018). Additionally, hierarchical structures also have been introduced because the context sentences do not have the same significance as the input sentences (Miculicich et al., 2018; Yun et al., 2020).

2.2 Coreference and NMT

The difference in coreference expressions among languages (Zinsmeister et al., 2017; Lapshinova-Koltunski et al., 2020) gives MT systems a challenge on pronoun translation (Bawden et al., 2018). Several recent works have attempted to incorporate coreference information (Ohtani et al., 2019). The closest work to ours is (Stojanovski and Fraser, 2018) which also adds noise on creating a coreference-augmented dataset, while we do not add oracle coreference information directly to the training data.

2.3 Data augmentation for NMT

One of the most common methods for data augmentation in NMT is back-translation that generates pseudo-parallel data from monolingual corpora using intermediate NMT models (Sennrich et al., 2016a). Generally, back-translation is conducted at sentence-level, however, several works have proposed document-level back-translation (Sugiyama and Yoshinaga, 2019; Huo et al., 2020).

On the other hand, sentence corruption by removing or replacing word(s) has also been widely used for improving model performance and robustness (Lample et al., 2018; Voita et al., 2019). Inspired by these works, we choose sentence corruption for contrastive learning.

2.4 Contrastive Learning

Contrastive learning is to learn a representation by contrasting positive and negative (contrastive) examples. It has succeed in various machine learning fields including computer vision (Chen et al., 2020) and natural language processing (Mikolov et al., 2013; Wu et al., 2020; Lee et al., 2021).

Recently, several approaches to contrastive learning for NMT have also been studied. Yang et al. (2019) proposed strategies for generating word-omitted contrastive examples and leveraging contrastive learning for reducing word omission errors in NMT. Pan et al. (2021) applied contrastive learning for multilingual MT and employed data augmentation for obtaining both the positive and negative training examples.

While these works have been conducted in sentence-level NMT settings, we focus on extending contrastive learning in context-aware NMT.

3 Context-aware NMT models

In this section, we briefly overview context-aware NMT methods and describe our baseline models which are also commonly adopted in recent works.

Generally, a sentence-level (context-agnostic) NMT model takes an input sentence in a source language and returns an output sentence in a target language. On the other hand, a context-aware NMT model is designed to handle surrounding contextual sentences of source and/or target sentences. We focus on leveraging the contextual sentences of the source language.

Throughout this work, we consider the Transformer (Vaswani et al., 2017) as a base model architecture by following the majority of the recent works on context-aware NMT. Transformer consists of a stack of self-attentional layers in which a self-attention module is followed by a feed-forward module for each layer. Here we list four Transformer-based configurations that we used in the experiments:

- **sent-level**: As a baseline, we have experimented with the basic Transformer model which does not use any contextual sentences.
- **concat**: This is a straightforward approach to incorporate contextual sentences without modifying the Transformer model (Tiedemann and Scherrer, 2017). This concatenates all contextual sentences and an input sentence with special tokens between sentences.
- **multi-enc**: This has an extra encoder for encoding contextual sentences separately. We follow the model introduced in (Voita et al., 2018) which obtain a hidden representation of contextual sentences by weight-shared Transformer encoder. The model combines the encoded source and context representations using a source-to-context attention mechanism and a gated summation.
- **multi-enc-hier**: To represent multiple contextual sentences effectively, hierarchical encoders for contextual sentences have been proposed (Miculicich et al., 2018; Yun et al., 2020). In this configuration, the context representation is calculated in token-level first, then finally processed in sentence-level. We experimented with the model of (Yun et al., 2020) in this paper.

All the model structures are described in Figure 1.

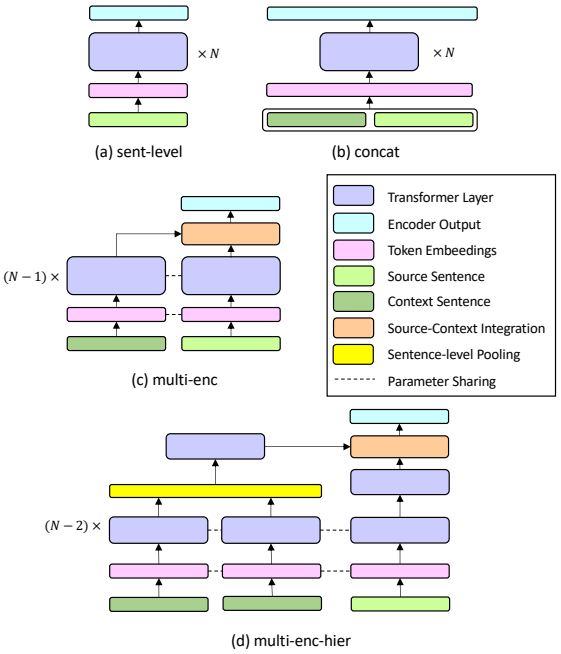

Figure 1: The structure of compared context-aware NMT models.

4 Our Method: CorefCL

In this section, we explain the main idea of CorefCL, a data augmentation and contrastive learning scheme leveraging coreference between the source and contextual sentences.

4.1 Data Augmentation Using Coreference

Generally, constrastive learning encourages a model to discriminate ground-truth and contrastive (negative) examples. In existing works, a number of approaches have been studied for obtaining contrastive examples:

- Corrupting the sentence by randomly removing or replacing one or more tokens in the sentence. (Yang et al., 2019)
- Choosing an irrelevant example in the batch or dataset. (Pan et al., 2021)
- Perturbations on representation space. Usually output vector of encoder or decoder is used. (Lee et al., 2021)

CorefCL basically takes a similar approach to the first one, by the sentence corruption. However, unlike previous works that modify the source sentence, CorefCL modifies the contextual sentences to form contrastive examples. Specifically, we corrupt cross-sentential coreference mentions which

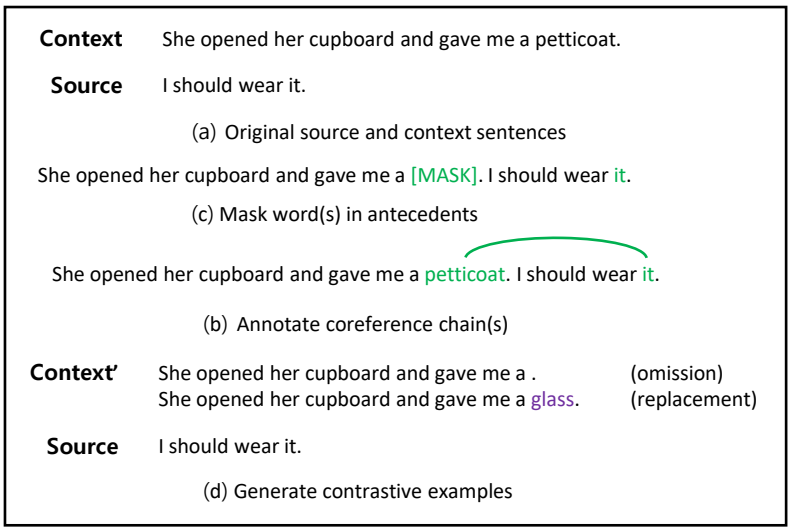

Figure 2: Data augmentation process of CorefCL.

occur between the source and its contextual sentences. This is based on the intuition that coreference is one of the core components of coherent translation.

More formally, steps to forming contrastive examples in CorefCL are as follows (see also Figure 2):

1. Annotate the source documents automatically. We use NeuralCoref[1] to identify the coreference mentions between the source and its previous sentences as contextual sentences

2. Filter the examples with cross-sentential coreference chain(s) between the source and contextual sentences. Around 20 to 30% of the training corpus is annotated in this way. See Section 5.1 for details

3. For each coreference chain, mask every word in the antecedents with a special token. We also keep the original examples for training

4. Masked words are replaced randomly with other words in vocabulary (*word replacement*), or omitted (*word omission*)

In the experiments, we take both of the corruption strategies. Precisely, the masked words are removed with a probability of 0.5, or randomly replaced otherwise. We found that this method is more effective compared to the methods using only one of the two corruption strategies. Please refer to the ablation study in Section 5.5 for more details.

[1]https://github.com/huggingface/neuralcoref

4.2 Contrastive Learning for Context-aware NMT

Context-aware NMT models can implicitly capture dependencies between the source and contextual sentences. CorefCL introduces a max-margin contrastive learning loss to train the model to explicitly discriminate inconsistent contexts. This contrastive loss also encourages a model to be more sensitive to the contents of contextual sentences.

Formally, given the source $\mathbf{x}$, target $\mathbf{y}$, n contextual sentences $C = [\mathbf{c}_1, \cdots, \mathbf{c}_n]$ in the data $\mathcal{D}$, we first train the model by minimizing a negative log-likelihood loss, which is a common MT loss:

$$\mathcal{L}_{MT} = \sum_{(\mathbf{x},\mathbf{y},C)\in\mathcal{D}} -\log P(\mathbf{y}|\mathbf{x}, C).$$

Once the model is trained with MT loss, we fine-tune the model with a contrastive loss. With a contrastive version of context $\tilde{C}$, our contrastive learning objective is minimizing a max-margin loss (Huang et al., 2018; Yang et al., 2019):

$$\mathcal{L}_{CL} = \sum_{(\mathbf{x},\mathbf{y},C,\tilde{C})\in\mathcal{D}} \max\{\eta + \log P(\mathbf{y}|\mathbf{x}, \tilde{C}) - \log P(\mathbf{y}|\mathbf{x}, C), 0\}.$$

Minimizing $\mathcal{L}_{CL}$ encourages the log-likelihood of the ground-truth to be at least η larger than that of the contrastive examples. In our formulation, we want the model to be more sensitive to the subtle changes in the contextual sentences.

The contrastive loss is jointly optimized with MT loss since we empirically found that the joint

optimization has yielded better performance than minimizing CL loss only as similar to (Yu et al., 2020):

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{MT} + \alpha\mathcal{L}_{CL},$$

where $\alpha \in [0, 1]$ is a weight for balancing between contrastive learning and MT loss. For simplicity, we fixed α during fine-tuning.

5 Experiments

5.1 Datasets

We experimented with CorefCL on various document-level parallel datasets: i) 3 English-German datasets including WMT document-level news translation[2] (Barrault et al., 2019), IWSLT TED talk [3] (Cettolo et al., 2017), OpenSubtitles'18[4] (Lison et al., 2018), and ii) our web-crawled English-Korean subtitles corpus.

For all tasks, we take every 2 preceding sentences as contextual sentences and we only consider sentences within the same document (article, talk, movie, one episode of TV programs, etc.) of the source sentence. If split of the validation and the test set is not presented in the data, we apply document-based split to ensure that training and validation/test data is well-separated. Details of datasets are listed as follows:

WMT We use a set of parallel corpora annotated with document boundaries which is released in WMT'19 news translation task. Specifically, we combine Europarl v9, News Commentary v14, and MODEL-RAPID to form a training set containing $3.7M$ examples and $0.85M$ with cross-sentential coreferences. For validation and test sets, we used newstest2013 and newstest2019 which contain $3.05k$ and $2.14k$ examples respectively.

IWSLT The IWSLT dataset consists of transcriptions of TED talks in a variety of languages. We used the 2017 version of the training set, a combination of dev2010, tst2010, tst2015 as a validation set, and tst2017 as a test set. The resulting dataset consists of $232k$ ($50.3k$ with cross-sentential coreferences), $3.5k$, $1.2k$ examples of train, dev, test sets respectively.

OpenSubtitles We also choose the English-German pair of OpenSubtitles2018 corpora. The raw corpus contains $24.4M$ parallel sentences. We follow the filtering methods in (Voita et al., 2019) by removing pairs that have a time overlap of subtitle frames less than 0.9. We also use separate documents for validation / test sets, resulting in $3.9M$ ($1.01M$ with cross-sentential coreferences), $40.7k$, $40.5k$ examples for train / validation / test sets respectively.

En-Ko Subtitles For English-Korean experiments, we first crawled approximately $6.1k$ bilingual subtitle files from websites such as GomLab.com. Since sentence pairs of these subtitles are already soft-aligned by the creators so we applied a simple time-code based heuristics to filter examples. The final data contains $1.6M$ ($0.24M$ with cross-sentential coreferences), $155.6k$, and $18.1k$ examples of consecutive sentences in the training, validation, and test sets respectively.

For preprocessing, all English and German corpus is tokenized first with Moses (Koehn et al., 2007) tokenizer[5]. We then apply the BPE (Sennrich et al., 2016b) using SentencePiece[6], and the size of the merge operation is approximately $16.5k$. We also put a special token [BOC] at the beginning of contextual sentences to differentiate them from the source sentences.

5.2 Settings

We use model hyperparameters, such as the size of hidden dimensions and the number of hidden layers as same the `transformer-base` (Vaswani et al., 2017), since all of the compared models are based on Transformer. Specifically, we set 512 as the hidden dimension, the number of layers is 6, the number of attention heads is 8, and the dropout rate is set to 0.1.

All models are trained with ADAM (Kingma and Ba, 2014) with different learning rates for each dataset. We employ early stopping of the training when the MT loss on the validation set does not improve. We start training each baseline model from scratch with random initialization and document-level dataset. Note that all the baseline models are not trained using iterative training as (Zhang et al., 2018; Huo et al., 2020) which first trains the model from sentence-level task first, then document-level task. All the evaluated models are implemented on top of the transformers[7] framework.

We measure the translation quality by the BLEU score (Papineni et al., 2002). For scoring BLEU,

[2] http://www.statmt.org/wmt19/translation-task.html
[3] https://wit3.fbk.eu/home
[4] https://opus.nlpl.eu/OpenSubtitles-v2018.php
[5] https://github.com/moses-smt/mosesdecoder
[6] https://github.com/google/sentencepiece
[7] https://github.com/huggingface/transformers

System	WMT	OpenSubtitles	IWSLT	En-Ko Subtitles	
				detok.	char.
sent-level	22.7	27.6	29.3	8.6	19.2
concat	22.4	28.3	29.7	9.3	22.1
+ CorefCL	23.5 (+1.1)	29.1 (+0.8)	30.9 (+1.3)	10.9 (+1.6)	24.9 (+2.8)
multi-enc	23.1	28.6	29.8	9.2	21.7
+ CorefCL	24.3 (+1.2)	29.8 (+1.4)	31.1 (+1.3)	10.8 (+1.6)	24.4 (+2.7)
multi-enc-hier	24.4	29.1	30.0	10.3	23.1
+ CorefCL	25.4 (+1.0)	30.2 (+1.1)	31.1 (+1.2)	11.7 (+1.4)	25.7 (+2.6)

Table 1: Corpus-level BLEU scores of compared models on different tasks. For the En-Ko subtitles task, we list both detokenized (detok.) and character-level (char.) scores. Improvements by CorefCL are denoted in (). Underlined score means that the model has the largest BLEU improvements among models in the same task.

we use the sacreBLEU (Post, 2018) case-sensitive, detokenized scores for En-De, and case-insensitive scores with `intl` tokenizer for En-Ko task. We also report case-insensitive char-level scores on En-Ko for comparison.

5.3 Overall BLEU Evaluation

We display the corpus-level test BLEU scores of all compared models on different tasks in Table 1. Among the baseline systems, all context-aware models show moderate improvements over the sentence-level (sent-level) baseline. These results are comparable to that of Huo et al. (2020) on the IWSLT task except for multi-enc-hier, and Yun et al. (2020) on OpenSubtitles task. One exception is a single-encoder model (concat) on WMT task, which seems due to the longer average sentence length.

We evaluated CorefCL by fine-tuning the context-aware models. Results show that models with CorefCL outperformed their vanilla counterparts, with the BLEU gain of up to 1.4 in En-De tasks, and 1.6/2.8 (detokenized/char-level BLEU) in the En-Ko subtitles task.

We observed that while CorefCL consistently improves BLEU on all tasks, it achieves better results on IWSLT and En-Ko subtitles tasks. Since improvements on much larger datasets like WMT and OpenSubtitles are smaller, we suggest that CorefCL also works as a regularization.

5.4 Results on English-German Contrastive Evaluation Set

To assess how CorefCL improves the ability to deal with pronoun-related translations more in detail, we experiment our method with ContraPro.[8] ContraPro is a contrastive test suit for En-De pronoun translation introduced by Müller et al. (2018). The evaluation is done by letting the model scores the German sentence with correct and incorrect pronoun translation, given the source and contextual English sentence. The accuracy is calculated by counting the number of correctly scored examples (i.e. correct examples that received a higher score than their incorrect counterpart).

System	Trained on			
	WMT		OpenSubtitles	
	BLEU	Acc.	BLEU	Acc.
sent-level	19.3	47.9	29.6	48.4
concat	19.9	49.7	30.5	54.4
+ CorefCL	20.3	51.2	32.3	57.9
multi-enc-hier	20.4	50.9	31.7	57.3
+ CorefCL	21.9	52.4	33.6	60.5

Table 2: BLEU and pronoun resolution accuracies on ContraPro (Müller et al., 2018) En-De contrastive test set.

We evaluate the models trained with WMT and OpenSubtitles tasks. We also list BLEU scores of En-De translation using the English source text in ContraPro. As shown in Table 2, CorefCL significantly improves the baselines in scoring accuracy for all models by up to 5.5%, as well as slight improvements in BLEU scores.

One interesting finding is that CorefCL also achieved substantial accuracy gain on the models trained on WMT. Since the ContraPro is created from OpenSubtitles, WMT-trained models would yield lower performance because of domain shift between training and testing. Table2 clearly shows the performance drop in BLEU, nevertheless, moderate improvements in accuracy can also be ob-

[8] https://github.com/ZurichNLP/ContraPro

served on WMT-trained models.

5.5 Analysis

System	BLEU	Accuracy
multi-enc-hier	31.7	57.3
+ CorefCL	33.6	60.5
- Word omission	32.4	59.4
- Word replacement	32.3	58.6

Table 3: Ablation study on coreference corruption strategy. All systems are trained on OpenSubtitles English-German dataset and evaluated on ContraPro.

Ablation Study CorefCL uses the two corruption strategies for generating contrastive coreference mentions; word omission and word replacement. To make a better understanding of influence of these strategies, we evaluate CorefCL of different settings of these strategies.

As shown in Table.3, using both types of corruptions results in better performance. Removing one of the two strategies slightly degrades both the pronoun resolution accuracy and BLEU. Although not being significant, removing the word replacement has more impact on accuracy. This suggests that a standard context-aware model, at least for multi-enc-hier is less sensitive to word substitution. The word replacement strategy can complement this behavior as resulted in better performance.

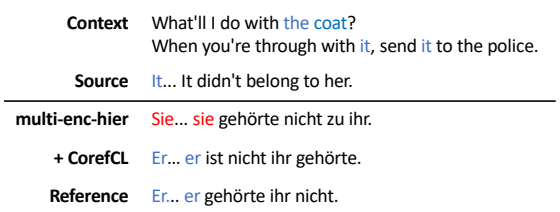

Figure 3: Example translation with and without CorefCL.

Qualitative Example We display a sample from ContraPro corpus and its translations made by multi-enc-hier model trained with OpenSubtitle task. In this example, since "coat" is translated as *Mantel* which is a masculine noun thus *Er* would be adequate translation of "It" instead of *Sie* which is feminine. While multi-enc-hier incorrectly translated "It" as *Sie*, the model fine-tuned with CorefCL correctly resolved it as *Er*.

In practice, context-aware models that do not leverage target-side contexts struggle to maintain these kinds of coreference consistency (Müller et al., 2018; Lapshinova-Koltunski et al., 2019) because of the asymmetric nature of grammatical components and data distributions. Results show that CorefCL can complement the limitation of source-only context-aware models.

6 Conclusions and Future Work

We have presented a data augmentation and contrastive learning scheme based on coreference for context-aware NMT. By leveraging coreference mentions between the source and target sentence, CorefCL effectively generates contrastive examples for applying contrastive learning on context-aware NMT models. In the experiments, CorefCL consistently improves the translation quality and pronoun resolution accuracy.

As future work, we plan to extend CorefCL to target contexts since maintaining coreference consistency needs both the source and the target contexts. It would be also interesting that applying CorefCL for fine-tuning pre-trained big language models like BART (Lewis et al., 2020) or T5 (Raffel et al., 2020) for downstream document-level MT tasks.

Acknowledgements

We thank Minwoo Lee for helpful discussions, as well as the anonymous reviewers for their thoughtful and constructive comments. This works is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2021R1A2C2008855)

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

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.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh,

Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 Evaluation Campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)*, pages 1–14.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. 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.

Jiaji Huang, Yi Li, Wei Ping, and Liang Huang. 2018. Large margin neural language model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191, Brussels, Belgium. Association for Computational Linguistics.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 604–616, Online. Association for Computational Linguistics.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint*, arXiv:1412.6980.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Cristina España-Bonet, and Josef van Genabith. 2019. Analysing coreference in transformer outputs. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 1–12, Hong Kong, China. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Marie-Pauline Krielke, and Christian Hardmeier. 2020. Coreference strategies in English-German translation. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 139–153, Barcelona, Spain (online). Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2021. Contrastive Learning with Adversarial Perturbations for Conditional Text Generation. In *ICLR*, pages 1–25.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence

learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

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.

Takumi Ohtani, Hidetaka Kamigaito, Masaaki Nagata, and Manabu Okumura. 2019. Context-aware neural machine translation with coreference information. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 45–50, Hong Kong, China. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online. 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Karin Sim Smith. 2017. On integrating discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121.

Dario Stojanovski and Alexander Fraser. 2018. Coreference and coherence in neural machine translation: A study using oracle experiments. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 49–60, Brussels, Belgium. Association for Computational Linguistics.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, December, pages 3104–3112, Montréal, Canada.

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? a targeted evaluation of neural machine translation architectures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4263–4272, Brussels, Belgium. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*.

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.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: Contrastive Learning for Sentence Representation. *arXiv preprint*, arXiv:2012.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Florence, Italy. Association for Computational Linguistics.

Lei Yu, Laurent Sartran, Po-Sen Huang, Wojciech Stokowiec, Domenic Donato, Srivatsan Srinivasan, Alek Andreev, Wang Ling, Sona Mokra, Agustin Dal Lago, Yotam Doron, Susannah Young, Phil Blunsom, and Chris Dyer. 2020. The DeepMind Chinese–English Document Translation System at WMT2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 326–337, Online. Association for Computational Linguistics.

Hyeongu Yun, Yongkeun Hwang, and Kyomin Jung. 2020. Improving context-aware neural machine translation using self-attentive sentence embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9498–9506.

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 the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Heike Zinsmeister, Stefanie Dipper, and Melanie Seiss. 2017. Abstract pronominal anaphors and label nouns in German and English: Selected case studies and quantitative investigations.