Reducing Word Omission Errors in Neural Machine Translation: A Contrastive Learning Approach

Zonghan Yang[♠] Yong Cheng[♣] Yang Liu[♠◇*] Maosong Sun[♠]

[♠]Institute for Artificial Intelligence
State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology
[♣]Google AI
[◇] Beijing Advanced Innovation Center for Language Resources

Abstract

While neural machine translation (NMT) has achieved remarkable success, NMT systems are prone to make word omission errors. In this work, we propose a contrastive learning approach to reducing word omission errors in NMT. The basic idea is to enable the NMT model to assign a higher probability to a ground-truth translation and a lower probability to an erroneous translation, which is automatically constructed from the ground-truth translation by omitting words. We design different types of negative examples depending on the number of omitted words, word frequency, and part of speech. Experiments on Chinese-to-English, German-to-English, and Russian-to-English translation tasks show that our approach is effective in reducing word omission errors and achieves better translation performance than three baseline methods.

1 Introduction

While neural machine translation (NMT) has achieved remarkable success (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017), there still remains a severe challenge: NMT systems are prone to omit essential words on the source side, which severely deteriorate the adequacy of machine translation. Due to the lack of interpretability of neural networks, it is hard to explain how these omission errors occur and design methods to eliminate them.

Existing methods for reducing word omission errors in NMT have focused on modeling coverage (Tu et al., 2016; Mi et al., 2016; Wu et al., 2016; Wang et al., 2016; Tu et al., 2017). The central idea is to model the fertility (i.e., the number of corresponding target words) of a source word based on attention weights to avoid word omission. Although these methods prove to be effective in modeling coverage for NMT, they heavily rely on the attention weights provided by the RNNsearch model (Bahdanau et al., 2015). Since the attention weights between input and output are not readily available in the state-of-the-art Transformer model (Vaswani et al., 2017), it is hard for existing methods to be directly applicable. As a result, it is important to develop model-agnostic methods for addressing the word omission problem in NMT.

In this paper, we propose a simple and effective *contrastive learning* approach to reducing word omission errors in NMT. The basic idea is to maximize the margin between the probability of a ground-truth translation and that of an erroneous translation for a given source sentence. The erroneous translations are automatically constructed via omitting words among the ground-truth translations. We design several types of erroneous translations in respect of omission counts, word frequency, and part of speech. Our approach has the following advantages:

- *Model agnostic*. Our approach is applicable to all existing NMT models. Only the training objective and training data need to be changed.
- *Language independent*. Our approach is independent of languages and can be applied to arbitrary languages.
- *Fast to train*. Contrastive learning starts with a pre-trained NMT model and usually converges in only hundreds of steps.

We evaluate our approach on German-to-English, Chinese-to-English, and Russian-to-English translation tasks. Experiments show that contrastive learning can not only effectively reduce word omission errors but also achieve better translation performance than existing methods in both automatic and human evaluations.

[*] Corresponding author: Yang Liu

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6191–6196
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

2 A Contrastive Learning Approach

Let $\mathbf{x}$ be a source sentence and $\mathbf{y}$ be a target sentence. We use $P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$ to denote an NMT model parameterized by $\boldsymbol{\theta}$. Given trained parameters $\hat{\boldsymbol{\theta}}$, the translation of a source sentence is given by

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y}} \Big\{ P(\mathbf{y}|\mathbf{x};\hat{\boldsymbol{\theta}}) \Big\} \quad (1)$$

During decoding process, the NMT model chooses the candidate sentence with the highest probability as the output translation. When a word omission error occurs, erroneous translations, which are mistakenly assigned with higher probabilities, are more likely to be chosen than ground-truth translations. Therefore, to reduce word omission errors, the probability that the NMT model assigns to an erroneous translation must be lower than that of a ground-truth translation.

Our proposed contrastive learning method is shown in Algorithm 1 , which consists of three steps. In the first step, the model is trained using maximum likelihood estimation (MLE) on a parallel corpus (lines 1-2). In the second step, negative examples are automatically constructed by omitting words in ground-truth translations (line 3). In the third step, the model is finetuned using contrastive learning with the estimates of MLE as a starting point.

More formally, given a parallel training set $D = \{\langle \mathbf{x}^{(s)}, \mathbf{y}^{(s)} \rangle\}_{s=1}^{S}$, the first step is to find a set of model parameters that maximizes the log-likelihood of the training set:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \Big\{ L(\boldsymbol{\theta}) \Big\}, \quad (2)$$

where the log-likelihood is defined as

$$L(\boldsymbol{\theta}) = \sum_{s=1}^{S} \log P(\mathbf{y}^{(s)}|\mathbf{x}^{(s)};\boldsymbol{\theta}) \quad (3)$$

The second step is to construct negative examples based on the ground-truth parallel corpus. Given a ground-truth sentence pair $\langle \mathbf{x}, \mathbf{y} \rangle$ from the parallel training set D, an erroneous sentence pair $\langle \mathbf{x}, \tilde{\mathbf{y}} \rangle$ can be automatically constructed by omitting words from the translation $\mathbf{y}$ in the ground-truth sentence pair. In this work, we distinguish between three methods for omitting words:

- *Random omission.* One or more source words are omitted according to a random uniform distribution.

Algorithm 1 Contrastive Learning for NMT

```
Input: $D = \{\langle \mathbf{x}^{(s)}, \mathbf{y}^{(s)} \rangle\}_{s=1}^{S}$
Output: $\hat{\boldsymbol{\theta}}_{\text{CL}}$
1: Obtain $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ using maximum likelihood estimation on $D$ with random initialization;
2: Construct $\tilde{D} = \{\langle \mathbf{x}^{(s)}, \tilde{\mathbf{y}}^{(s)} \rangle\}_{s=1}^{S}$ based on $D$ automatically;
3: Obtain $\hat{\boldsymbol{\theta}}_{\text{CL}}$ using contrastive learning on $\tilde{D}$ with $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ as a starting point.
```

- *Omission by word frequency.* One or more source words are omitted according to word frequencies.
- *Omission by part of speech.* One or more source words are omitted according to parts of speech.

Contrastive learning starts with the model parameters trained by MLE. Our contrastive learning approach is equipped with a max-margin loss. The max-margin loss ensures that the margins of the log-likelihood between the ground-truth pairs and the contrastive examples are higher than the setting η:

$$\hat{\boldsymbol{\theta}}_{\text{CL}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \Big\{ J(\boldsymbol{\theta}) \Big\}, \quad (4)$$

where the max-margin loss is defined as

$$J(\boldsymbol{\theta}) = \sum_{s=1}^{S} \max \Bigg\{ \sum_{n=1}^{N} \log P(\tilde{\mathbf{y}}_n^{(s)}|\mathbf{x}^{(s)};\boldsymbol{\theta}) + \eta - N \log P(\mathbf{y}^{(s)}|\mathbf{x}^{(s)};\boldsymbol{\theta}), 0 \Bigg\}. \quad (5)$$

For each ground-truth sentence pair $\langle \mathbf{x}^{(s)}, \mathbf{y}^{(s)} \rangle$, it is possible to sample N negative examples $\langle \mathbf{x}^{(s)}, \tilde{\mathbf{y}}_1^{(s)} \rangle, \dots, \langle \mathbf{x}^{(s)}, \tilde{\mathbf{y}}_N^{(s)} \rangle$. For simplicity, we set $N = 1$ and use $\tilde{D} = \{\langle \mathbf{x}^{(s)}, \tilde{\mathbf{y}}^{(s)} \rangle\}_{s=1}^{S}$ in our experiments.

3 Experiments

We evaluated the proposed method on Chinese-to-English, German-to-English, and Russian-to-English translation tasks.

3.1 Setup

For the Chinese-to-English translation task, we use the WMT 2017 dataset as the training set, which is composed of the News Commentary v12, UN Parallel Corpus v1.0, and CWMT corpora. The training set contains 25M sentence pairs. The newsdev2017 and newstest2017 datasets are used as the development set and test set, respectively. For the German-to-English translation task, we use the WMT 2017 dataset as the training set, which consists of 6M preprocessed sentence pairs. The newstest2014 and newstest2017 datasets are used as the development set and test set, respectively. For the Russian-to-English translation task, we use the WMT 2017 preprocessed dataset as the training set, which consists of 25M preprocessed sentence pairs. The newstest2015 and newstest2016 datasets are used as the development set and test set, respectively.

Following Sennrich et al. (2016b), we split words into sub-word units. The numbers of merge operations in byte pair encoding (BPE) for both language pairs are set to 32K. After performing BPE, the training set of the Chinese-to-English task contains 550M Chinese sub-word units and 615M English sub-word units, the training set of the German-to-English task consists of 157M German sub-word units and 153M English sub-word units, and the training set of the Russian-to-English task consists of 653M Russian sub-word units and 629M English sub-word units.

We used three baselines in our experiments:

- MLE: Maximum likelihood estimation. The setting of hyper-parameters is the same with (Vaswani et al., 2017);
- MLE + CP: Imposing the coverage penalty (Wu et al., 2016) constraint on the decoding process of MLE. We treat the softmax weight matrix in the uppermost "encoder-decoder attention" layer of Transformer as the attention weight matrix to calculate coverage penalty;
- WordDropout: Implementing the word dropout technique proposed by Sennrich et al. (2016a) during MLE training.

For our contrastive learning method, we compare different settings of erroneous training set $\tilde{D}$:

- $\text{CL}_{\text{one/two/three}}$: $\tilde{D}$ is constructed via omitting one/two/three words randomly from the ground-truth translations in D;
- $\text{CL}_{\text{low/high}}$: $\tilde{D}$ is constructed via omitting the word with the lowest/highest frequency from each ground-truth translation in D;

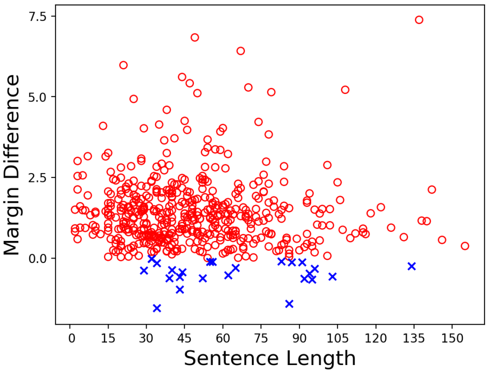

Figure 1: Visualization of margin differences between CL_{one} and MLE on 500 sampled sentence pairs. We use red to highlight sentence pairs on which CL_{one} achieves a larger margin than MLE. Blue points denote MLE achieves a higher margin.

- $\text{CL}_{\text{V/IN}}$: $\tilde{D}$ is constructed via omitting one verb or preposition randomly from the ground-truth translation in D. The part-of-speech information is given by the Stanford Parser (Manning et al., 2014).

3.2 Comparison of Margins

To find out whether CL increases the margin compared with MLE, we calculate the following margin difference for a ground-truth sentence pair $\langle \mathbf{x}, \mathbf{y} \rangle$ and an erroneous sentence pair $\langle \mathbf{x}, \tilde{\mathbf{y}} \rangle$:

$$\Delta M = \log P(\mathbf{y}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{\text{CL}}) - \log P(\tilde{\mathbf{y}}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{\text{CL}}) - \log P(\mathbf{y}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{\text{MLE}}) + \log P(\tilde{\mathbf{y}}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{\text{MLE}}) \quad (6)$$

Figure 1 shows the margin difference between CL_{one} and MLE on 500 sampled sentence pairs from the training set for the Chinese-to-English task. "Sentence length" denotes the sum of the lengths of the source and target sentences (i.e., $|\mathbf{x}| + |\mathbf{y}|$). Red points denote sentence pairs on which CL_{one} has a larger margin than MLE (i.e., $\Delta M > 0$), while the blue ones denote the $\Delta M < 0$ case. We find that CL_{one} has a larger margin than MLE on 95% of the 500 sampled sentence pairs, with an average margin difference of 1.4.

3.3 Automatic Evaluation Results

Table 1 shows the results of automatic evaluation on Chinese-to-English, German-to-English, and Russian-to-English translation tasks. The evaluation metric is case-insensitive BLEU score (Papineni et al., 2002). Contrastive learning starts with the model parameters trained by MLE and converges in only 150 steps. For fair comparison, all

Method	Zh-En	De-En	Ru-En
MLE	23.90	34.88	31.24
MLE + CP	24.04	34.93	31.36
WordDropout	23.73	34.63	31.05
$\mathrm{CL}_{\mathrm{one}}$	$24.92^{++**\dagger\dagger}$	$35.74^{++**\dagger\dagger}$	$32.04^{++**\dagger\dagger}$
$\mathrm{CL}_{\mathrm{two}}$	$24.76^{++**\dagger\dagger}$	$35.54^{++**\dagger\dagger}$	$31.94^{++*\dagger\dagger}$
$\mathrm{CL}_{\mathrm{three}}$	$24.52^{+*\dagger\dagger}$	$35.44^{++*\dagger\dagger}$	$32.20^{++**\dagger\dagger}$
$\mathrm{CL}_{\mathrm{low}}$	$24.13^{\dagger}$	$34.96^{\dagger}$	$31.47^{++\dagger}$
$\mathrm{CL}_{\mathrm{high}}$	$24.77^{++**\dagger\dagger}$	$35.24^{++\dagger\dagger}$	$31.70^{++\dagger\dagger}$
CL_{V}	$24.12^{\dagger}$	$35.02^{\dagger\dagger}$	$31.73^{++*\dagger\dagger}$
$\mathrm{CL}_{\mathrm{IN}}$	$24.71^{++**\dagger\dagger}$	$35.26^{+*\dagger\dagger}$	$31.76^{++*\dagger\dagger}$

Table 1: Automatic evaluation results on Chinese-to-English, German-to-English, and Russian-to-English translation tasks. Contrastive learning starts with the model parameters trained by MLE and converges in only 150 steps. For fair comparison, all the models of MLE, MLE + CP, and MLE + data are trained for another 150 steps as well, but yielding no further improvement. "+": significantly better than MLE ($p < 0.05$). "++": significantly better than MLE ($p < 0.01$). "$*$": significantly better than MLE + CP ($p < 0.05$). "$**$": significantly better than MLE + CP ($p < 0.01$)."†": significantly better than WordDropout ($p < 0.05$). "††": significantly better than WordDropout ($p < 0.01$).

	Method	Flu.	Ade.
Evaluator 1	MLE	4.31	4.25
	MLE + CP	4.31	4.31
	WordDropout	4.29	4.25
	$\mathrm{CL}_{\mathrm{one}}$	4.32	4.58
Evaluator 2	MLE	4.27	4.22
	MLE + CP	4.26	4.25
	WordDropout	4.25	4.23
	$\mathrm{CL}_{\mathrm{one}}$	4.27	4.53

Table 2: Human evaluation results on the Chinese-to-English task. "Flu." denotes fluency and "Ade." denotes adequacy. Two human evaluators who can read both Chinese and English were asked to assess the fluency and adequacy of the translations. The scores of fluency and adequacy range from 1 to 5.

Method	Zh-En	De-En	Ru-En
MLE	362	221	471
MLE + CP	265	200	383
WordDropout	245	168	351
$\mathrm{CL}_{\mathrm{one}}$	122	138	250

Table 3: Comparison of error counts on the test sets. CL denotes the contrastive learning method with the highest BLEU score, which is $\mathrm{CL}_{\mathrm{one}}$ for the Chinese-to-English and German-to-English tasks and $\mathrm{CL}_{\mathrm{three}}$ for the Russian-to-English task.

the models of MLE, MLE+CP, and MLE+data are trained for another 150 steps as well, but yielding no further improvement.

We observe that with negative examples synthesized properly, our contrastive learning method significantly outperforms MLE, MLE + CP, and WordDropout on all three language pairs.

An interesting finding is that omitting high-frequency source words (i.e., $\mathrm{CL}_{\mathrm{high}}$) achieves significantly better results than omitting low-frequency source words (i.e., $\mathrm{CL}_{\mathrm{low}}$) for all three language pairs, which suggests that standard NMT models tend to omit high-frequency source words rather than low-frequency words.

The experiment on omission by part of speech further confirms this finding as omitting high-frequency prepositions (i.e., $\mathrm{CL}_{\mathrm{IN}}$) leads to better results than omitting low-frequency verbs (i.e., CL_{V}).

3.4 Human Evaluation Results

Table 2 shows the results of human evaluation on the Chinese-to-English task. We asked two human evaluators who can read both Chinese and English to evaluate the fluency and adequacy of the translations generated by MLE, MLE + CP, MLE + data, and $\mathrm{CL}_{\mathrm{one}}$. The scores of fluency and adequacy range from 1 to 5. The translations were shuffled randomly, and the name of each method was anonymous to human evaluators.

We find that CL_{one} significantly improves the adequacy over all baselines. This is because omitting important information in source sentences decreases the adequacy of translation. CL_{one} is capable of alleviating this problem by assigning lower probabilities to translations with word omission errors.

To further quantify to what extent our approach reduces word omission errors, we asked human evaluators to manually count word omission errors on the test sets of all the translation tasks. Table 3 shows the error counts. We find that $\mathrm{CL}_{\mathrm{one}}$ achieves significant error reduction as compared with MLE, $\mathrm{MLE} + \mathrm{CP}$, and $\mathrm{WordDropout}$ for all the three language pairs.

4 Related Work

Our work is related to two lines of research: modeling coverage for NMT and contrastive learning in NLP.

4.1 Modeling Coverage for NMT

The notion of coverage dates back to conventional phrase-based statistical machine translation (Koehn et al., 2003). A coverage vector, which is used to indicate whether a source phrase is translated or not during the decoding process, ensures that each source phrase is translated exactly once. As there are no latent variables defined on language structures in neural networks, it is hard to directly introduce coverage into NMT. As a result, there are two strategies. The first strategy is to modify the model architectures to incorporate coverage (Tu et al., 2016; Mi et al., 2016), which requires considerable expertise. The second strategy is to impose constraints on the decoding process (Wu et al., 2016).

Our work differs from prior studies in that contrastive learning is model agnostic. All previous coverage-based methods heavily rely on attention weights between source and target words to derive coverage for source words. Such attention weights are not readily available for all NMT models. In contrast, our method can be used to fine-tune arbitrary NMT models to reduce word omission errors in only hundreds of steps.

4.2 Contrastive Learning in NLP

Contrastive learning has been widely used in natural language processing. For instance, word embeddings are usually learned by the noise contrastive estimation method (Gutmann and Hyvärinen, 2012): a negative example is synthesized by randomly selecting a word from the vocabulary to replace a word in a ground-truth example (Vaswani et al., 2013; Mnih and Kavukcuoglu, 2013; Bose et al., 2018).

Contrastive learning has also been investigated in neural language modelling (Huang et al., 2018), unsupervised word alignment (Liu and Sun, 2015), order embeddings (Vendrov et al., 2016; Bose et al., 2018), knowledge graph embeddings (Yang et al., 2015; Lin et al., 2015; Bose et al., 2018) and caption generation (Mao et al., 2016; Vedantam et al., 2017).

The closest work to ours is (Wiseman and Rush, 2016), which leverages contrastive learning during beam search with the golden reference sentences as positive examples and the current output sentences as contrastive examples. While they focus on improving the capability of Seq2Seq model to capture global dependencies, we focus on reducing word omission errors of Transformer model effectively.

5 Conclusion

We have presented contrastive learning for reducing word omission errors in neural machine translation. Contrastive examples are automatically constructed by omitting words from the ground-truth translations. Our approach is model-agnostic and can be applied to arbitrary NMT models. Experiments show that our approach significantly reduces omission errors and improves translation performance on three language pairs.

6 Acknowledgments

We thank all anonymous reviewers for their valuable comments. This work is supported by the National Key R&D Program of China (No. 2017YFB0202204), National Natural Science Foundation of China (No. 61761166008, No. 61432013), Beijing Advanced Innovation Center for Language Resources (No. TYR17002), and the NExT++ project supported by the National Research Foundation, Prime Ministers Office, Singapore under its IRC@Singapore Funding Initiative.

References

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Avishek Joey Bose, Huan Ling, and Yanshuai Cao.

2018. Adversarial contrastive estimation. In *Proceedings of ACL*.

Michael U Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(Feb):307–361.

Jiaji Huang, Yi Li, Wei Ping, and Liang Huang. 2018. Large margin neural language model. In *Proceedings of EMNLP*.

Phillip Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL*.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of AAAI*.

Yang Liu and Maosong Sun. 2015. Contrastive unsupervised word alignment with non-local features. In *Proceedings of AAAI*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of CVPR*.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. In *Proceedings of EMNLP*.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Proceedings of NIPS*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of ACL*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *Proceedings of AAAI*.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of ACL*.

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 *Proceedings of NIPS*.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with large-scale neural language models improves translation. In *Proceedings of EMNLP*.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of CVPR*.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *Proceedings of ICLR*.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016. A novel approach to dropped pronoun translation. In *Proceedings of NAACL*.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. In *Proceedings of NIPS*.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of ICLR*.