Enhancing Neural Machine Translation with Syntactic Ambiguities

Anonymous ACL submission

Abstract

Benefiting from the data-driven end-to-end model architecture, neural machine translation has obvious performance advantages over statistical machine translation, but its demand for data is also significantly greater, including monolingual and parallel corpus. Most of the past studies have focused on reducing the demand for parallel corpus or making more effective use of limited parallel corpus. In this work, we have studied a method of using ambiguity of syntactic structure to achieve more effective use of monolingual corpus. Experiments conducted on multiple benchmarks for various languages show that our method has a greater improvement than the method using back-translation only, demonstrating the effectiveness of our proposed method.

1 Introduction

The end-to-end neural machine translation (NMT) model could achieve good translation results only by relying on parallel corpus without other manually designed features (Bahdanau et al., 2015; Vaswani et al., 2017). A typical NMT model is an encoder-decoder architecture, where the encoder is responsible for encoding the source language input, and the decoder generates the target language translation according to the source language representation. Therefore, parallel corpus is needed to train the encoder-decoder model during the training stage, and usually the more high-quality parallel corpus, the better the translation effect of the trained model.

In machine translation, monolingual corpus is often used to enhance the translation performance. In the era of statistical machine translation (SMT), starting from the IBM model (Brown et al., 1990), monolingual target sentences are used to improve the fluency of translations, such as using language models in phrase SMT systems (Koehn et al., 2003; Brants et al., 2007).

NMT systems can also benefit from language models trained on monolingual corpus (He et al., 2016; Gülçehre et al., 2017; Domhan and Hieber, 2017). Besides, monolingual corpus is also commonly used in unsupervised or semi-supervised NMT training settings. On the one hand, the NMT model can be pre-trained on monolingual corpus (Conneau and Lample, 2019; Song et al., 2019). Pre-training methods on monolingual corpus usually include denoising and masked language modeling. The former method adds noise to the sentence as input and then requires the model to restore the original sentence, and the latter method requires the model to predict the masked tokens of the input with the remaining ones. On the other hand, the pseudo-parallel corpus can be synthesized for translation training, i.e., back-translation (Sennrich et al., 2016a; Poncelas et al., 2018; Edunov et al., 2018; Caswell et al., 2019).

In back-translation, to make the most use of the monolingual text, Imamura et al. (2018) show that sampling synthetic sources is more effective than beam search, thus resulting multiple sources for each target. Whereas Edunov et al. (2018) perform sampling or noised beam strategies on only a single sample, opting to train on a larger number of target sentences instead. Hoang et al. (2018); Cotterell and Kreutzer (2018) propose an iterative procedure which continuously produce different pseudo-parallel pairs to improve the final translation quality. Different from these existing works, our work starts from the perspective of ambiguity in language structure and uses ambiguity to generate different sentence versions, thereby generating different translations, thus forming more pseudo-parallel sentence pairs, and ultimately improving the performance of the NMT system.

We evaluated our method on five classical benchmarks: WMT14 En→De, En→Fr, Fr→En, WMT17 De→En and WMT20 En→Zh. Compared our method with back-translation and sampling+back-translation baselines, we have a significant performance improvement. Our contribution is that we used syntactic ambiguity in machine translation for the first time to improve translation performance. The proposed method is simple and easy to use, without the need to increase the amount of monolingual data, which is meaningful for some scenarios with limited parallel and monolingual data.

2 Method

2.1 Syntactic Ambiguity

Syntactic ambiguity in natural language processing can be defined as a phenomenon that a sentence is structurally ambiguous when it can be assigned to more than one syntactic structure (Zavrel et al., 1997). The resolution of syntactic structural ambiguity is one of the central problems in natural language analysis. Figure 1 shows two syntactic structures of the sentence "*President Bush called his attention with this method*". Both syntactic structures are valid, and different syntactic structures will bring about different syntactic meanings. In Figure 1(a) structure is the PP "*with this method*" is attached to the verb "*called*", while in Figure 1(b), the PP "*with this method*" does not attach to the verb but to the NP "*his attention*". This structural ambiguity shown in Figure 1 is called Prepositional Phrase (PP) attachment, which is the drosophila of structural ambiguity resolution.

This type of ambiguity is very common in some languages, such as English, German, French, and Chinese, where there is very little overt case marking and syntactic information alone does not suffice to explain the difference in attachment sites between such sentences. For natural language understanding, it is necessary to use semantic and even pragmatic information to re-analyze sentences in order to make correct decisions (Hindle and Rooth, 1991). But we do the opposite, and use the changes in sentence meaning brought about by this ambiguity to construct more single sentences and more to dig out the role of limited corpus.

2.2 Enhancement in Back-translation

Back-translation has been shown to be an effective method for improving the performance of machine translation models using monolingual data. Formally, for languages S and T in back-translation, given parallel corpus $D^P = \langle D_S^P, D_T^P \rangle$, monolingual corpus D_S^M, D_T^M, first train the initial $T \rightarrow S$ translation model $\mathcal{M}_{T\rightarrow S}$ based on D^P. Second, use the translation model $\mathcal{M}_{T\rightarrow S}$ to translate D_T^M into language S to get $\hat{D}_S^M$, thus forming pseudo-parallel corpus pairs $\langle \hat{D}_S^M, D_T^M \rangle$ with D_T^M. Third, combine the synthesized pseudo-parallel corpus $\langle \hat{D}_S^M, D_T^M \rangle$ with the original parallel corpus D^P to obtain a new mixed parallel corpus for training the translation direction $S \rightarrow T$ translation model $\mathcal{M}_{S\rightarrow T}$.

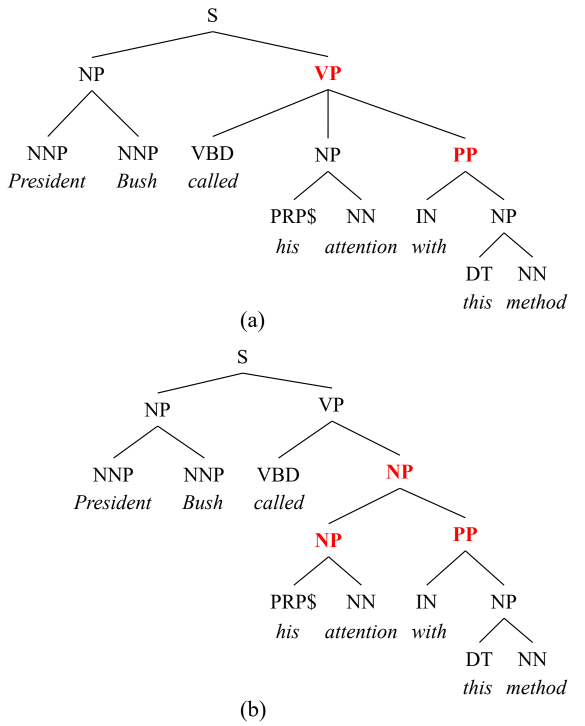

Figure 1: An example of syntactic ambiguity for sentence *President Bush called his attention with this method*.

Iterative back-translation can be used to further improve performance if bi-directional monolingual data is available. Specifically, the training process includes N iteration steps. For each step, first use the pseudo-parallel corpus obtained in the previous step $\langle \hat{D}_S^M, D_T^M \rangle$ and $\langle \hat{D}_T^M, D_S^M \rangle$ to combine the parallel corpus D^P to train $S \rightarrow T$ and $T \rightarrow S$ translation models $\mathcal{M}_{S\rightarrow T}$ and $\mathcal{M}_{T\rightarrow S}$ respectively. And then use the new obtained $\mathcal{M}_{S\rightarrow T}$ and $\mathcal{M}_{T\rightarrow S}$ to translate the monolingual sentences D_S^M and D_T^M to $\hat{D}_T^M$ and $\hat{D}_S^M$, forming a new pseudo-parallel corpus $\langle \hat{D}_S^M, D_T^M \rangle$ and $\langle \hat{D}_T^M, D_S^M \rangle$, which are used for the next training. For the first step, since there is no pseudo-parallel corpus, only the parallel corpus is used to train the model.

We use syntactic ambiguity to construct different meaning versions of the same sentence through explicit structural declarations. We define this construction process as $G(\cdot)$. Through the amplification of monolingual sentences with $G(\cdot)$, more pseudo-parallel corpus will be generated during the back-translation training process, thereby enhancing back-translation.

For the sentence amplification process $G(\cdot)$, since we need to be able to explicitly control the meaning of the sentence to remove the ambiguity and get its definite meaning version, we refer to the rules in mathematical operations and use parentheses to control the priority of PP attachment, so as to obtain different deterministic grammar structure. Specifically, we use a simple and effective search algorithm (as shown in Algorithm 1) on the constituent syntax parse tree, insert parentheses to different positions for obtaining the final sentence sequences with different meanings. It is worth noting that the Chinese PP constituent is preceded, so the algorithm is to find the next sibling, rather than looking for the previous one as in English.

Algorithm 1: Amplification Process $G(\cdot)$

```
Input: Constituent parse tree T of sentence s;
U = {s};
for t ∈ T do
    if t.label == PP then
        for st ∈ t.parent do
            if st is the previous sibling of t then
                b = st.start;
                e = t.end;
                s_c = InsertParentheses(s, b, e);
                U = U ∪ {s_c};
                b = st.start;
                e = st.end;
                s_c = InsertParentheses(s, b, e);
                U = U ∪ {s_c};
InsertParentheses(s, b, e)
    return s[: b] ⊙ "(" ⊙ s[b : e] ⊙ ")" ⊙ s[e :];
Output: U.
```

Take "*President Bush called his attention with this method*" as an example, after the amplified process, the sentence becomes a set {*President Bush called his attention with this method*, *President Bush called (his attention) with this method*, *President Bush called (his attention with this method)*}", Using the backward translation model to translate into Chinese: "{布什总统用这种方法引起了他的注意，布什总统用这种方法引起了（他的注意），布什总统呼吁（他用这种方法注意）}". Then we remove the added parentheses and duplicated sentences to get the final pseudo-parallel sentence pairs: {⟨ 布什总统用这种方法引起了他的注意, *President Bush called his attention with this method* ⟩, ⟨布什总统呼吁他用这种方法注意, *President Bush called his attention with this method* ⟩}. Our enhancement method can be used for normal back-translation with only monolingual data in the target language, or iterative back-translation with monolingual data in both languages.

3 Experiments

3.1 Setup

We conducted a series of experiments on the classic machine translation benchmarks to verify the effectiveness of our proposed method, including WMT14 En→De, En→Fr, Fr→En, WMT17 De→En and WMT20 En→Zh. Among them, De→En, Fr→En are to verify the effectiveness of the proposed method in English, while En→De, En→Fr, En→Zh are to verify the universality of the method in more languages. We train our model on all available bitext using the official settings, excluding sentences longer than 250 words and sentence pairs with a source/target length ratio greater than 1.5. We sampled 10M sentences for each language from newscrawl monolingual data.

Following the common practice, we tokenize all sentences with the Moses tokenizer (Koehn et al., 2007) except Chinese and learn a joint source and target Byte-Pair-Encoding (BPE) (Sennrich et al., 2016b) with 40K types. For Chinese sentences, we employed the Jieba[1] morphological analyzer to segment the sentences into words. With the exception of En→Zh, we report the majority of our results in terms of case-sensitive tokenized BLEU (Papineni et al., 2002), but we also report de-tokenized BLEU scores using sacreBLEU (Post, 2018). We provide a character-level BLEU score for En→Zh evaluation. For model configuration, follow the practice of (Vaswani et al., 2017), we use the *transformer.big* setting with embedding dimension / FFN layer dimension / number of layers 1024 / 4096 / 6 respectively. Label smoothing (Szegedy et al., 2016; Pereyra et al., 2017) with a uniform prior distribution over the vocabulary $\epsilon = 0.1$ is employed for all models.

3.2 Results and Analysis

We show the evaluation results of WMT14 En→De, En→Fr, Fr→En, WMT17 De→En in Table 1. From the results in the table, back-translation has a

[1] https://github.com/fxsjy/jieba

Model	WMT14 En→De		WMT14 En→Fr		WMT14 Fr→En		WMT17 De→En	
	BLEU	sacreBLEU	BLEU	sacreBLEU	BLEU	sacreBLEU	BLEU	sacreBLEU
Baseline	28.45	27.3	41.20	39.3	28.75	27.1	32.35	31.5
+back-translation								
greedy	29.70	28.4	42.35	40.2	29.88	28.9	33.91	32.7
beam	29.55	28.1	42.02	40.0	29.54	28.3	33.84	32.5
noise beam	30.86	29.1	42.94	41.0	31.07	30.3	34.35	33.2
sampling	31.65	29.8	43.26	41.3	31.52	30.6	34.52	33.5
ambiguity	31.68	29.8	43.19	41.1	31.68	30.6	34.60	33.6
sampling+ambiguity	**32.16**	**30.1**	**43.89**	**41.6**	**32.05**	**30.9**	**35.05**	**33.9**
+iterative back-translation								
greedy	30.31	28.7	42.89	40.9	31.67	30.3	34.34	33.3
sampling	32.08	30.0	43.76	41.4	32.60	31.2	34.92	34.0
ambiguity	32.20	30.0	43.69	41.4	32.59	31.3	34.95	34.0
sampling+ambiguity	**32.97**	**30.5**	**44.23**	**41.9**	**33.56**	**32.6**	**35.60**	**34.7**

Table 1: Results on WMT14 En→De, En→Fr, Fr→En and WMT17 De→En test sets. Results shown in bold are better than the corresponding baselines at significance level $p < 0.01$ (Collins et al., 2005).

Model	BLEU	Δ
Baseline	38.75	–
+back-translation		
greedy	39.54	0.79 ↑
sampling	40.32	1.57 ↑
ambiguity	40.41	1.66 ↑
sampling+ambiguity	**41.06**	2.31 ↑
+iterative back-translation		
greedy	40.15	1.40 ↑
sampling	41.08	2.33 ↑
ambiguity	40.95	2.20 ↑
sampling+ambiguity	**41.54**	2.79 ↑

Table 2: Results on WMT20 En→Zh test set.

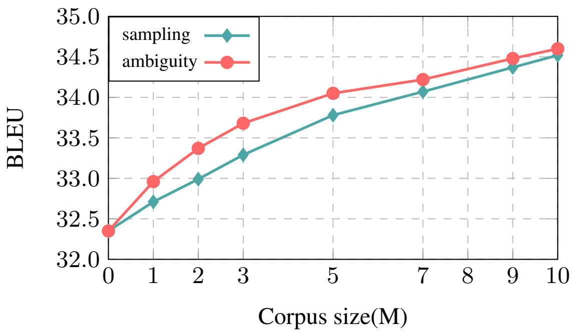

Figure 2: The impact of synthetic pseudo-parallel corpus size on WMT17 De→En translation performance.

large performance improvement compared to the baseline, and iterative back-translation is improved more significantly, which shows that the target monolingual can effectively improve the model performance through back-translation and the monolingual at both ends can further improves by simultaneously helping the forward and backward translation model to get better at the same time. sampling and noise beam strategies are better than greedy and beam in back-translation, which shows that increasing the diversity of generation can effectively improve the effect of back-translation.

Our back-translation based on the ambiguity strategy achieves a similar enhancing effect as the sampling strategy, but the contribution of our method is orthogonal to the sampling method, and we have obtained better translation effects by further superimposing these two strategies. The translation effect of WMT20 En→Zh shown in Table 2 also shows a similar phenomenon. And the results on En→De, En→Fr, En→Zh show that syntax ambiguity can not only be used in English, but also adaptable in other languages.

We further explored the effect of ambiguity and sampling strategies under different monolingual scales in Figure 2. As shown in the figure, our ambiguity strategy is more effective when the monolingual scale is relatively small.

4 Conclusion

In this work, we change the back-translation input from the perspective of the ambiguity of the syntactic structure rather than sampling the model prediction probability distribution for synthesizing more pseudo-parallel pairs to achieve the purpose of enhancement. We have conducted experiments on multiple machine translation benchmarks, and the results show that our method can improve both back-translation and iterative back-translation baseline. And our method can also cooperate with sampling, which utilize the uncertainty of prediction for enhancement, to play a stronger effect.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic. Association for Computational Linguistics.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Ryan Cotterell and Julia Kreutzer. 2018. Explaining and generalizing back-translation through wake-sleep. *CoRR*, abs/1806.04402.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Comput. Speech Lang.*, 45:137–148.

Wei He, Zhongjun He, Hua Wu, and Haifeng Wang. 2016. Improved neural machine translation with SMT features. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 151–157. AAAI Press.

Donald Hindle and Mats Rooth. 1991. Structural ambiguity and lexical relations. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 229–236, Berkeley, California, USA. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Kenji Imamura, Atsushi Fujita, and Eiichiro Sumita. 2018. Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 55–63, Melbourne, Australia. Association for Computational Linguistics.

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.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

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.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Alberto Poncelas, Dimitar Sht. Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman

Passban. 2018. Investigating backtranslation in neural machine translation. *CoRR*, abs/1804.06189.

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.

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.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

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 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jakub Zavrel, Walter Daelemans, and Jorn Veenstra. 1997. Resolving PP attachment ambiguities with memory-based learning. In *CoNLL97: Computational Natural Language Learning*.