Agreement-based Joint Training for Bidirectional Attention-based Neural Machine Translation

Yong Cheng#, Shiqi Shen†, Zhongjun He+, Wei He+, Hua Wu†, Maosong Sun†, Yang Liu†∗

#Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
†State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, China
+Baidu Inc., Beijing, China

Abstract
The attentional mechanism has proven to be effective in improving end-to-end neural machine translation. However, due to the structural divergence between natural languages, unidirectional attention-based models might only capture partial aspects of attentional regularities. We propose agreement-based joint training for bidirectional attention-based end-to-end neural machine translation. Instead of training source-to-target and target-to-source translation models independently, our approach encourages the two complementary models to agree on word alignment matrices on the same training data. Experiments on Chinese-English and English-French translation tasks show that joint training significantly improves both alignment and translation quality over independent training.

1 Introduction
End-to-end neural machine translation (NMT) is a newly proposed paradigm for machine translation [Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015]. Without explicitly modeling latent structures that are vital for conventional statistical machine translation [Brown et al., 1993; Koehn et al., 2003; Chiang, 2005], NMT builds on an encoder-decoder framework: the encoder transforms a source-language sentence into a continuous-space representation, from which the decoder generates a target-language sentence.

While early NMT models encode a source sentence as a fixed-length vector, Bahdanau et al. [2015] advocate the use of attention in NMT. They indicate that only parts of the source sentence have an effect on the target word being generated. In addition, the relevant parts often vary with different target words. Such an attentional mechanism has proven to be an effective technique in text generation tasks such as machine translation [Bahdanau et al., 2015; Luong et al., 2015] and image caption generation [Xu et al., 2015].

However, due to the structural divergence between natural languages, modeling the correspondence between words in two languages still remains a major challenge for NMT, especially for distantly-related languages. For example, Luong et al. [2015] report that attention-based NMT lags behind the Berkeley
aligner [Liang et al., 2006] in terms of alignment error rate (AER) on the English-German data. One possible reason is that unidirectional attention-based NMT can only capture partial aspects of attentional regularities due to the non-isomorphism of natural languages.

In this work, we propose to introduce agreement-based learning [Liang et al., 2006, 2007] into attention-based neural machine translation. The basic idea is to encourage source-to-target and target-to-source translation models to agree on word alignment on the same training data. This can be done by defining a new training objective that combines likelihoods in two directions as well as an agreement term that measures the consensus between word alignment matrices in two directions. Experiments on Chinese-English and English-French datasets show that our approach achieves significant improvements in terms of alignment and translation quality as compared with independent training.

2 Background

Given a source-language sentence $x = x_1, \ldots, x_m, \ldots, x_M$ that contains $M$ words and a target-language sentence $y = y_1, \ldots, y_n, \ldots, y_N$ that contains $N$ words, end-to-end neural machine translation directly models the translation probability as a single, large neural network:

$$ P(y|x; \theta) = \prod_{n=1}^{N} P(y_n|x, y_{<n}; \theta) $$

(1)

where $\theta$ is a set of model parameters and $y_{<n} = y_1, \ldots, y_{n-1}$ is a partial translation.
The encoder-decoder framework [Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015] usually uses a recurrent neural network (RNN) to encode the source sentence into a sequence of hidden states $h = h_1, \ldots, h_m, \ldots, h_M$:

$$h_m = f(x_m, h_{m-1}, \theta)$$  \hspace{1cm} (2)

where $h_m$ is the hidden state of the $m$-th source word and $f$ is a non-linear function. Note that there are many ways to obtain the hidden states. For example, Bahdanau et al. [2015] use a bidirectional RNN and concatenate the forward and backward states as the hidden state of a source word to capture both forward and backward contexts. Figure 1 illustrates how the decoder generates the first target word $y_1$ and the target hidden state $s_1$ given the concatenation of forward and backward source hidden states.

Bahdanau et al. [2015] define the conditional probability in Eq. (1) as

$$P(y_n|x, y_{<n}; \theta) = g(y_{n-1}, s_n, c_n, \theta)$$ \hspace{1cm} (3)

where $g$ is a non-linear function and $s_n$ is the hidden state corresponding to the $n$-th target word computed by

$$s_n = f(s_{n-1}, y_{n-1}, c_n, \theta)$$ \hspace{1cm} (4)

The context vector $c_n$ for generating the $n$-th target word is calculated as

$$c_n = \sum_{m=1}^{M} A(\theta)_{n,m} h_m$$ \hspace{1cm} (5)

where $A(\theta)$ is an alignment matrix, in which an element $A(\theta)_{n,m}$ reflects the contribution of the $m$-th source word $x_m$ to generating the $n$-th target word $y_n$:

$$A(\theta)_{n,m} = \frac{\exp(a(s_{n-1}, h_m, \theta))}{\sum_{m'=1}^{M} \exp(a(s_{n-1}, h_{m'}, \theta))}$$ \hspace{1cm} (6)

where $a(s_{n-1}, h_m, \theta)$ measures how well $x_m$ and $y_n$ are aligned. Note that word alignment is treated as a function parametrized by $\theta$ instead of a latent variable in attention-based NMT.

Given a set of training examples $\{(x^{(s)}, y^{(s)})\}_{s=1}^{S}$, the training algorithm aims to find the model parameters that maximize the likelihood of the training data:

$$\theta^* = \arg\max_{\theta} \left\{ \sum_{s=1}^{S} \log P(y^{(s)}|x^{(s)}; \theta) \right\}$$ \hspace{1cm} (7)

Although the introduction of attention has advanced the state-of-the-art of NMT, it is still challenging for attention-based NMT to capture the intricate structural divergence between natural languages. Figure 2(a) shows the Chinese-to-English (upper) and English-to-Chinese (bottom) alignment matrices for the same sentence pair. Both the two independently trained models fail to correctly capture the gold-standard correspondence: while the Chinese-to-English alignment assigns wrong probabilities to “us” and “bush”, the English-to-Chinese alignment makes wrong predictions on “condemns” and “bombing”.

Fortunately, although each model only captures partial aspects of the mapping between words in natural languages, the two models seem to be complementary: the Chinese-to-English alignment does well on “condemns” and the English-to-Chinese alignment assigns correct probabilities to “us” and “bush”. Therefore, combining the two models can hopefully improve both alignment and translation quality in both directions.
3 Agreement-based Joint Training

In this work, we propose to introduce agreement-based learning [Liang et al., 2006, 2007] into attention-based neural machine translation. The central idea is to encourage the source-to-target and target-to-source models to agree on alignment matrices on the same training data.

More formally, we train both the source-to-target attention-based neural translation model \( P(y|x; \theta) \) and the target-to-source model \( P(x|y; \theta) \) on a set of training examples \( \{ (x^{(s)}, y^{(s)}) \}_{s=1}^{S} \), where \( \theta \) and
\( \theta \) are model parameters in two directions, respectively. The new training objective is given by

\[
J(\theta, \bar{\theta}) = \sum_{s=1}^{S} \log P(x^{(s)} | y^{(s)}; \theta) + \sum_{s=1}^{S} \log P(y^{(s)} | x^{(s)}; \bar{\theta}) - \lambda \sum_{s=1}^{S} \Delta(\tilde{A}^{(s)}(\theta), \tilde{A}^{(s)}(\bar{\theta}))
\]

(8)

where \( \tilde{A}^{(s)}(\theta) \) is the source-to-target alignment matrix for the \( s \)-th sentence pair, \( \tilde{A}^{(s)}(\bar{\theta}) \) is the target-to-source alignment matrix for the \( s \)-th sentence pair, \( \Delta(\cdot) \) is a loss function that measures the discrepancy between two matrices, and \( \lambda \) is a hyper-parameter that balances the preference between likelihoods and agreement.

Inspired by the agreement term \[\text{Liang et al., 2006}\] and model invertibility regularization \[\text{Levinboim et al., 2015}\], we use the following loss function in our experiments:

\[
\Delta(\tilde{A}^{(s)}(\theta), \tilde{A}^{(s)}(\bar{\theta})) = -\log \sum_{n=1}^{N} \sum_{m=1}^{M} \tilde{A}^{(s)}(\theta)_{n,m} \times \tilde{A}^{(s)}(\bar{\theta})_{m,n}
\]

(9)

Intuitively, this loss function encourages both models to agree on each cell in the alignment matrices. As shown in Figure 2, joint learning leads to increased consensus between source-to-target and target-to-source models.

Our goal is to find the model parameters that maximize the training objective:

\[
\hat{\theta}^* = \arg\max_{\theta} \left\{ \sum_{s=1}^{S} \log P(x^{(s)} | y^{(s)}; \theta) - \lambda \sum_{s=1}^{S} \Delta(\tilde{A}^{(s)}(\theta), \tilde{A}^{(s)}(\bar{\theta})) \right\}
\]

(10)

\[
\hat{\bar{\theta}}^* = \arg\max_{\bar{\theta}} \left\{ \sum_{s=1}^{S} \log P(y^{(s)} | x^{(s)}; \bar{\theta}) - \lambda \sum_{s=1}^{S} \Delta(\tilde{A}^{(s)}(\theta), \tilde{A}^{(s)}(\bar{\theta})) \right\}
\]

(11)

It is easy to use the stochastic gradient descent (SGD) algorithm to implement agreement-based joint learning since the translation models in two directions share the same training data.

4 Experiments

4.1 Setup

We evaluated our approach on Chinese-English and English-French datasets.

For Chinese-English, the training corpus from LDC consists of 2.56M sentence pairs with 67.53M Chinese words and 74.81M English words. We used the NIST 2006 dataset as the validation set for hyper-parameter optimization and model selection and the NIST 2002, 2003, 2004, 2005, and 2008 datasets as the test sets. In the NIST Chinese-English datasets, each Chinese sentence has four corresponding English translations. To build English-Chinese evaluation datasets, we select the first English sentence in the four references as the source sentence and the Chinese sentence as the single reference translation.

For English-French, the training corpus from WMT 2014 consists of 12.07M sentence pairs with 303.88M English words and 348.24M French words. We follow Bahdanau et al. [2015] to restrict that sentences are no longer than 50 words. The concatenation of news-test-2012 and news-test-2013 is used as the validation set and news-test-2014 as the test set. The French-English evaluation sets can be easily obtained by reversing the English-French datasets.

We compared our approach with two state-of-the-art SMT and NMT systems.
Table 1: Results on the Chinese-English translation task. Moses is a phrase-based statistical machine translation system [Koehn and Hoang, 2007]. GroundHog is an attention-based neural machine translation system [Bahdanau et al., 2015]. We introduce agreement-based joint training for bidirectional attention-based NMT. NIST06 is the validation set and NIST02-05, 08 are test sets. The BLEU scores are case-insensitive. “*”: significantly better than Moses ($p < 0.05$); “**”: significantly better than Moses ($p < 0.01$); “+”: significantly better than GroundHog with independent training ($p < 0.05$); “++”: significantly better than GroundHog with independent training ($p < 0.01$).

| System    | Train. | Direct. | NIST06 | NIST02 | NIST03 | NIST04 | NIST05 | NIST08 |
|-----------|--------|---------|--------|--------|--------|--------|--------|--------|
| Moses     | Indep. | C→E     | 32.48  | 32.69  | 32.39  | 33.62  | 30.23  | 25.17  |
|           |        | E→C     | 14.27  | 18.28  | 15.36  | 13.96  | 14.11  | 10.84  |
| GroundHog | Indep. | C→E     | 30.74  | 35.16  | 33.75  | 34.63  | 31.74  | 23.63  |
|           |        | E→C     | 15.71  | 20.76  | 16.56  | 16.85  | 15.14  | 12.70  |
|           | Joint  | C→E     | 32.65  | 35.68††| 34.79†††| 35.72‡‡‡| 32.98‡‡‡| 25.62‡‡|
|           |        | E→C     | 16.25†††| 21.70‡‡‡| 17.45‡‡‡| 16.98‡‡| 15.70‡‡| 13.80‡‡‡|

Table 2: Results on the Chinese-English word alignment task. The evaluation metric is alignment error rate. “++”: significantly better than GroundHog with independent training ($p < 0.01$).

| training | | C→E | E→C |
|----------|----------|-----|-----|
| indep.   | 54.64    | 52.49 |
| joint    | 47.49**  | 46.70** |

1. Moses [Koehn and Hoang, 2007]: a phrase-based SMT system;
2. GroundHog [Bahdanau et al., 2015]: an attention-based NMT system.

For Moses, we used the parallel corpus to train the phrase-based translation model and the target-side part of the parallel corpus to train a 4-gram language model using the SRILM toolkit [Stolcke, 2002]. For GroundHog, we use the parallel corpus to train the neural machine translation models. The vocabulary size is set to 30K for all languages. Our approach simply extends GroundHog by replacing independent training with agreement-based joint training. The encoder-decoder framework and the attentional mechanism remain unchanged. We follow Jean et al. [2015] to address unknown words based on alignment matrices. Given alignment matrices between source sentences and target sentences, it is possible to calculate the position of the source word that is most translationally equivalent for each target word. After a source sentence is translated, each unknown word is translated from its corresponding source word. While Jean et al. [2015] use bilingual dictionary generated by an off-the-shelf word aligner to translate unknown words, we simply use unigram phrases.

4.2 Results on Chinese-English Data

Table 1 shows the results on the Chinese-English translation task. We find that our approach significantly outperforms both Moses and GroundHog with independent training in both Chinese-to-English and English-to-Chinese directions across all datasets.

Figure 2(b) shows example alignment matrices resulted from agreement-based joint training. We find that agreement-based joint training improves the alignment accuracy significantly since the two models are complementary.
Table 3: Results on the English-French translation task. The BLEU scores are case-insensitive. "***": significantly better than Moses ($p < 0.01$); "++": significantly better than GroundHog with independent training ($p < 0.01$).

Table 2 shows the results on the Chinese-English word alignment task. We used the TsinghuaAligner evaluation dataset [Liu and Sun, 2015] in which both the validation and test sets contain 450 manually-aligned Chinese-English sentence pairs. We follow Luong et al. [2015] to "force" decode our jointly trained models to produce translations that match the references. Then, we extract only one-to-one alignments by selecting the source word with the highest alignment weight for each target word. We find that agreement-based joint training significantly reduces alignment errors for both directions as compared with independent training. 1

4.3 Results on English-French Data

Table 3 gives the results on the English-French translation task. While GroundHog with independent training achieves translation performance on par with Moses, agreement-based joint learning leads to significant improvements over both baselines. This suggests that our approach is general and can be applied to more language pairs.

5 Related Work

Our work is inspired by two lines of research: (1) attention-based neural machine translation and (2) agreement-based learning.

5.1 Attention-based Neural Machine Translation

Bahdanau et al. [2015] first introduce the attentional mechanism into neural machine translation to enable the decoder to focus on relevant parts of the source sentence during decoding. The attention mechanism allows a neural model to cope better with long sentences because it does not need to encode all the information of a source sentence into a fixed-length vector regardless of its length. In addition, the attentional mechanism allows us to look into the “black box” to gain insights on how NMT works from a linguistic perspective.

1The error rates in Table 2 are still higher than conventional aligners that can achieve an AER of 30. One reason is that we only extract one-to-one word alignments while most gold-standard alignments for Chinese-English are usually one-to-many or many-to-many. In addition, there is no empty cept [Brown et al., 1993] in attention-based NMT to handle unaligned words.
Luong et al. [2015] propose two simple and effective attentional mechanisms for neural machine translation and compare various alignment functions. They show that attention-based NMT are superior to non-attentional models in translating names and long sentences.

After analyzing the alignment matrices generated by GROUNDHOG [Bahdanau et al., 2015], we find that modeling the structural divergence of natural languages is so challenging that unidirectional models can only capture part of alignment regularities. This finding inspires us to improve attention-based NMT by combining two unidirectional models. In this work, we only apply agreement-based joint learning to GROUNDHOG. As our approach does not assume specific network architectures, it is possible to apply it to the models proposed by Luong et al. [2015].

5.2 Agreement-based Learning

Liang et al. [2006] first introduce agreement-based learning into word alignment. The basic idea is to encourage asymmetric IBM models to agree on word alignment, which is a latent structure in word-based translation models [Brown et al., 1993]. This strategy significantly improves alignment quality across many languages. They extend this idea to deal with more latent-variable models in grammar induction and predicting missing nucleotides in DNA sequences [Liang et al., 2007].

Liu et al. [2015] propose generalized agreement for word alignment. The new general framework allows for arbitrary loss functions that measure the disagreement between asymmetric alignments. The loss functions can not only be defined between asymmetric alignments but also between alignments and other latent structures such as phrase segmentations.

In attention-based NMT, word alignment is treated as a parametrized function instead of a latent-variable. This makes word alignment differentiable, which is important for training attention-based NMT models. Although alignment matrices in attention-based NMT are in principle “symmetric” as they allow for many-to-many soft alignments, we find that unidirectional modeling can only capture partial aspects of structure mapping. Our contribution is to adapt agreement-based learning into attentional NMT, which leads to significant improvements in terms of both alignment and translation performance.

6 Conclusion

We have presented an approach to agreement-based joint training for bidirectional attention-based neural machine translation. By encouraging source-to-target and target-to-source models to agree on parametrized alignment matrices, joint learning achieves significant improvements in terms of alignment and translation quality over independent training. In the future, we plan to further validate the effectiveness of our approach on more language pairs.

acknowledgement

This work is done when Yong Cheng and Shiqi Shen visited Baidu as interns. This research is supported by the 973 Program (2011CBA00300, 2011CBA00301, 2014CB340501, 2014CB340505) and the National Natural Science Foundation of China (No. 61033001, 61361136003, 61522204).

References

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

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 1993.

David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL*, 2005.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of EMNLP*, 2014.

Sebastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On using very large target vocabulary for neural machine translation. In *Proceedings of ACL*, 2015.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of EMNLP*, 2013.

Philipp Koehn and Hieu Hoang. Factored translation models. In *Proceedings of EMNLP*, 2007.

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

Tomer Levinboim, Ashish Vaswani, and David Chiang. Model invertibility regularization: Sequence alignment with or without parallel data. In *Proceedings of NAACL*, 2015.

Percy Liang, Ben Taskar, and Dan Klein. Alignment by agreement. In *Proceedings of NAACL*, 2006.

Percy Liang, Dan Klein, and Michael I. Jordan. Agreement-based learning. In *Proceedings of NIPS*, 2007.

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

Chunyang Liu, Yang Liu, Huanbo Luan, Maosong Sun, and Heng Yu. Generalized agreement for bidirectional word alignment. In *Proceedings of EMNLP*, 2015.

Minh-Thanh Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*, 2015.

Andreas Stolcke. Srilm - an extensible language modeling toolkit. In *Proceedings of ICSLP*, 2002.

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

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, KyungHyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of ICML*, 2015.