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 intricate 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 agreement-based 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 (SMT) [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, Bahadanau 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.*, 2015b] and image caption generation [Xu *et al.*, 2015].

*Yang Liu is the corresponding author: liuyang2011@ tsinghua.edu.cn.

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. [2015b] 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 is capable of better accounting for attentional regularities and significantly improves alignment and translation quality over independent training.

2 Background

Given a source-language sentence $\mathbf{x} = \mathbf{x}_1, \ldots, \mathbf{x}_m, \ldots, \mathbf{x}_M$ that contains M words and a target-language sentence $\mathbf{y} = \mathbf{y}_1, \ldots, \mathbf{y}_n, \ldots, \mathbf{y}_N$ that contains N words, end-to-end neural machine translation directly models the translation probability as a single, large neural network:

$$P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \prod_{n=1}^{N} P(\mathbf{y}_n|\mathbf{x},\mathbf{y}_{<n};\boldsymbol{\theta}) \quad (1)$$

where $\boldsymbol{\theta}$ is a set of model parameters and $\mathbf{y}_{<n} = \mathbf{y}_1, \ldots, \mathbf{y}_{n-1}$ 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 $\mathbf{h} = \mathbf{h}_1, \ldots, \mathbf{h}_m, \ldots, \mathbf{h}_M$:

$$\mathbf{h}_m = f(\mathbf{x}_m, \mathbf{h}_{m-1}, \boldsymbol{\theta}) \quad (2)$$

where $\mathbf{h}_m$ is the hidden state of the m-th source word and $f(\cdot)$ is a non-linear function. Note that there are many ways

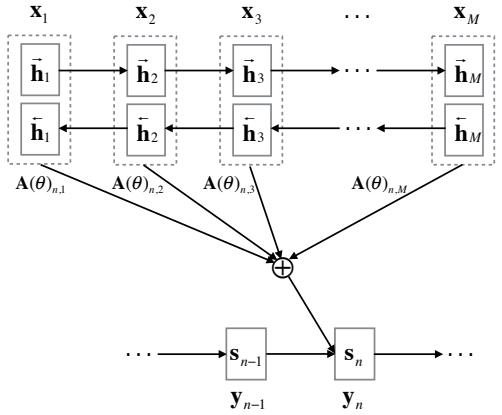


Figure 1: The illustration of attention-based NMT. The decoder generates a target hidden state $\mathbf{s}_n$ and its corresponding target word $\mathbf{y}_n$ given a source sentence $\mathbf{x}$. A bidirectional RNN is used to concatenate the forward and backward states as the hidden states of source words.

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 (see Figure 1).

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

$$P(\mathbf{y}_n|\mathbf{x}, \mathbf{y}_{<n}; \boldsymbol{\theta}) = g(\mathbf{y}_{n-1}, \mathbf{s}_n, \mathbf{c}_n, \boldsymbol{\theta}) \tag{3}$$

where $g(\cdot)$ is a non-linear function, $\mathbf{s}_n$ is the hidden state corresponding to the n-th target word computed by

$$\mathbf{s}_n = f(\mathbf{s}_{n-1}, \mathbf{y}_{n-1}, \mathbf{c}_n, \boldsymbol{\theta}) \tag{4}$$

and $\mathbf{c}_n$ is a context vector for generating the n-th target word:

$$\mathbf{c}_n = \sum_{m=1}^{M} \mathbf{A}(\boldsymbol{\theta})_{n,m} \mathbf{h}_m \tag{5}$$

We refer to $\mathbf{A}(\boldsymbol{\theta}) \in \mathbb{R}^{N \times M}$ as *alignment matrix*, in which an element $\mathbf{A}(\boldsymbol{\theta})_{n,m}$ reflects the contribution of the m-th source word $\mathbf{x}_m$ to generating the n-th target word $\mathbf{y}_n$: [1]

$$\mathbf{A}(\boldsymbol{\theta})_{n,m} = \frac{\exp(a(\mathbf{s}_{n-1}, \mathbf{h}_m, \boldsymbol{\theta}))}{\sum_{m'=1}^{M} \exp(a(\mathbf{s}_{n-1}, \mathbf{h}_{m'}, \boldsymbol{\theta}))} \tag{6}$$

where $a(\mathbf{s}_{n-1}, \mathbf{h}_m, \boldsymbol{\theta})$ measures how well $\mathbf{x}_m$ and $\mathbf{y}_n$ are aligned. Note that word alignment is treated as a function parameterized by $\boldsymbol{\theta}$ instead of a latent variable in attention-based NMT.

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

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

[1] We denote the alignment matrix as $\mathbf{A}(\boldsymbol{\theta})$ instead of α in [Bahdanau *et al.*, 2015] to emphasize that it is a function parameterized by $\boldsymbol{\theta}$ and differentiable. Although $\mathbf{s}_n$ and $\mathbf{c}_n$ also depend on $\boldsymbol{\theta}$, we omit the dependencies for simplicity.

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 incorrect 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 alignment and translation quality in both directions.

3 Agreement-based Joint Training

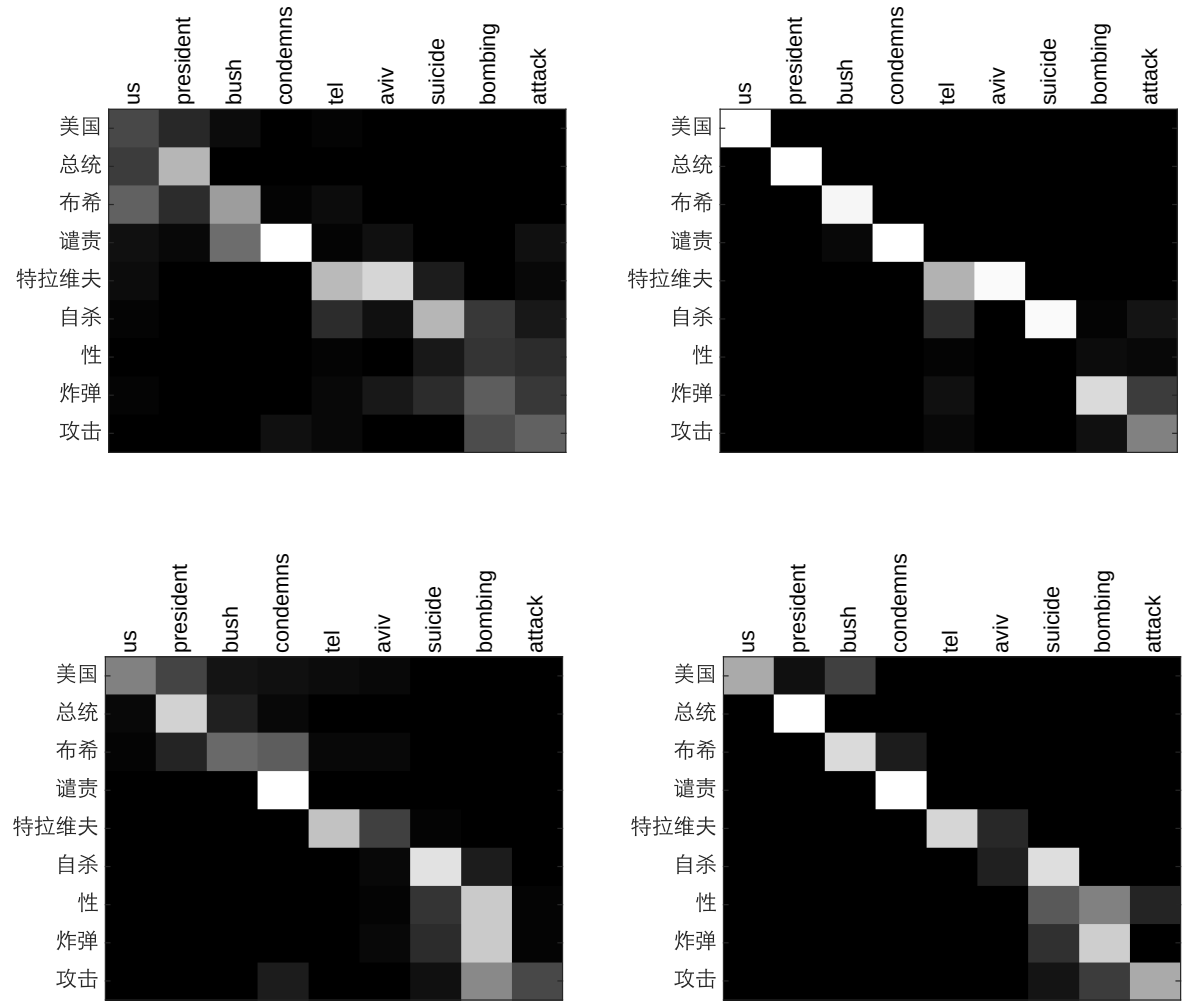
In this work, we propose to introduce agreement-based learning [Liang *et al.*, 2006; 2007] into attention-based neural machine translation. The central idea is to encourage the source-to-target and target-to-source models to agree on alignment matrices on the same training data. As shown in Figure 2(b), agreement-based joint training is capable of removing unlikely attention and resulting in more concentrated and accurate alignment matrices in both directions.

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

$$\begin{aligned} J(\overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}}) &= \sum_{s=1}^{S} \log P(\mathbf{y}^{(s)}|\mathbf{x}^{(s)}; \overrightarrow{\boldsymbol{\theta}}) \\ &+ \sum_{s=1}^{S} \log P(\mathbf{x}^{(s)}|\mathbf{y}^{(s)}; \overleftarrow{\boldsymbol{\theta}}) \\ &- \lambda \sum_{s=1}^{S} \Delta\Big(\mathbf{x}^{(s)}, \mathbf{y}^{(s)}, \overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\Big) \end{aligned} \tag{8}$$

where $\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}})$ is the source-to-target alignment matrix for the s-th sentence pair, $\overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})$ is the target-to-source alignment matrix for the same sentence pair, $\Delta(\cdot)$ is a loss function that measures the disagreement between two matrices, and λ is a hyper-parameter that balances the preference between likelihood and agreement.

For simplicity, we omit the dependency on the sentence pair and simply write the loss function as $\Delta\big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\big)$. While there are many alternatives for quantifying disagreement, we use the following three types of loss functions in our experiments:



(a) independent training

(b) joint training

Figure 2: Example alignments of (a) independent training and (b) joint training on a Chinese-English sentence pair. The first row shows Chinese-to-English alignments and the second row shows English-to-Chinese alignments. We find that the two unidirectional models are complementary and encouraging agreement leads to improved alignment accuracy.

1. *Square of addition* (SOA): the square of the element-wise addition of corresponding matrix cells

$$\Delta_{\text{SOA}}\big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\big) = -\sum_{n=1}^{N}\sum_{m=1}^{M}\Big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}})_{n,m} + \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})_{m,n}\Big)^2 \quad (9)$$

Intuitively, this loss function encourages to increase the sum of the alignment probabilities in two corresponding matrix cells.

2. *Square of subtraction* (SOS): the square of the element-wise subtraction of corresponding matrix cells

$$\Delta_{\text{SOS}}\big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\big) = \sum_{n=1}^{N}\sum_{m=1}^{M}\Big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}})_{n,m} - \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})_{m,n}\Big)^2 \quad (10)$$

Derived from the symmetry constraint proposed by Ganchev et al. [2010], this loss function encourages that an aligned pair of words share close or even equal alignment probabilities in both directions.

3. *Multiplication* (MUL): the element-wise multiplication of corresponding matrix cells

$$\Delta_{\text{MUL}}\big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\big) = -\log\sum_{n=1}^{N}\sum_{m=1}^{M}\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}})_{n,m} \times \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})_{m,n} \quad (11)$$

This loss function is inspired by the agreement term

[Liang *et al.*, 2006] and model invertibility regularization [Levinboim *et al.*, 2015].

The decision rules for the two directions are given by

$$\overrightarrow{\boldsymbol{\theta}}^* = \operatorname*{argmax}_{\overrightarrow{\boldsymbol{\theta}}} \left\{ \sum_{s=1}^{S} \log P(\mathbf{y}^{(s)}|\mathbf{x}^{(s)}; \overrightarrow{\boldsymbol{\theta}}) - \lambda \sum_{s=1}^{S} \Delta\big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\big) \right\} \quad (12)$$

$$\overleftarrow{\boldsymbol{\theta}}^* = \operatorname*{argmax}_{\overleftarrow{\boldsymbol{\theta}}} \left\{ \sum_{s=1}^{S} \log P(\mathbf{x}^{(s)}|\mathbf{y}^{(s)}; \overleftarrow{\boldsymbol{\theta}}) - \lambda \sum_{s=1}^{S} \Delta\big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\big) \right\} \quad (13)$$

Note that all the loss functions are differentiable with respect to model parameters. It is easy to extend the original training algorithm for attention-based NMT [Bahdanau *et al.*, 2015] to implement agreement-based joint training since the two 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 machine translation tasks.

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. The NIST 2002, 2003, 2004, 2005, and 2008 datasets were used as test sets. In the NIST Chinese-English datasets, each Chinese sentence has four reference English translations. To build English-Chinese validation and test sets, we simply "reverse" the Chinese-English datasets: 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. The concatenation of news-test-2012 and news-test-2013 was used as the validation set and news-test-2014 as the test set. Each English sentence has a single reference French translation. 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:

1. MOSES [Koehn and Hoang, 2007]: a phrase-based SMT system;
2. RNNSEARCH [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 [Stolcke, 2002]. We used the default system setting for both training and decoding.

For RNNSEARCH, we used the parallel corpus to train the attention-based NMT models. The vocabulary size is set to 30K for all languages. We follow Jean et al. [2015] to address the unknown word problem based on alignment matrices. Given an alignment matrix, it is possible to calculate the position of the source word to which is most likely to be aligned 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 a bilingual dictionary generated by an off-the-shelf word aligner to translate unknown words, we use unigram phrases instead.

Our system simply extends RNNSEARCH by replacing independent training with agreement-based joint training. The encoder-decoder framework and the attentional mechanism remain unchanged. The hyper-parameter λ that balances the preference between likelihood and agreement is set to 1.0 for Chinese-English and 2.0 for English-French. The training time of joint training is about 1.2 times longer than that of independent training for two directional models. We used the same unknown word post-processing technique as RNNSEARCH for our system.

Loss	BLEU
Δ_{SOA}: square of addition	31.26
Δ_{SOS}: square of subtraction	31.65
Δ_{MUL}: multiplication	32.65

Table 1: Comparison of loss functions in terms of case-insensitive BLEU scores on the validation set for Chinese-to-English translation.

4.2 Comparison of Loss Functions

We first compared the three loss functions as described in Section 3 on the validation set for Chinese-to-English translation. The evaluation metric is case-insensitive BLEU.

As shown in Table 1, the square of addition loss function (i.e., Δ_{SOA}) achieves the lowest BLEU among the three loss functions. This can be possibly attributed to the fact that a larger sum does not necessarily lead to increased agreement. For example, while 0.9 + 0.1 hardly agree, 0.2 + 0.2 perfectly does. Therefore, Δ_{SOA} seems to be an inaccurate measure of agreement.

The square of subtraction loss function (i.e, Δ_{SOS}) is capable of addressing the above problem by encouraging the training algorithm to minimize the difference between two probabilities: $(0.2 - 0.2)^2 = 0$. However, the loss function fails to distinguish between $(0.9 - 0.9)^2$ and $(0.2 - 0.2)^2$. Apparently, the former should be preferred because both models have high confidence in the matrix cell. It is unfavorable for two models agree on a matrix cell but both have very low confidence. Therefore, Δ_{SOS} is perfect for measuring agreement but ignores confidence.

As the multiplication loss function (i.e., Δ_{MUL}) is able to take both agreement and confidence into account (e.g., $0.9 \times 0.9 > 0.2 \times 0.2$), it achieves significant improvements over Δ_{SOA} and Δ_{SOS}. As a result, we use Δ_{MUL} in the following experiments.

System	Training	Direction	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
RNNSEARCH	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 translation task. MOSES is a phrase-based statistical machine translation system. RNNSEARCH is an attention-based neural machine translation system. 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 RNNSEARCH with independent training ($p < 0.05$); "++": significantly better than RNNSEARCH with independent training ($p < 0.01$). We use the statistical significance test with paired bootstrap resampling [Koehn, 2004].

Training	C → E	E → C
indep.	54.64	52.49
joint	47.49**	46.70**

Table 3: Results on the Chinese-English word alignment task. The evaluation metric is alignment error rate. "**": significantly better than RNNSEARCH with independent training ($p < 0.01$).

4.3 Results on Chinese-English Translation

Table 2 shows the results on the Chinese-to-English (C → E) and English-to-Chinese (E → C) translation tasks. [2] We find that RNNSEARCH generally outperforms MOSES except for the C → E direction on the NIST08 test set, which confirms the effectiveness of attention-based NMT on distantly-related language pairs such as Chinese and English.

Agreement-based joint training further systematically improves the translation quality in both directions over independently training except for the E → C direction on the NIST04 test set.

4.4 Results on Chinese-English Alignment

Table 3 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. [2015b] 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. This suggests that introducing agreement does enable NMT to capture attention more accurately and thus lead to better translations. Figure 2(b) shows example alignment matrices resulted from agreement-based joint training.

However, the error rates in Table 3 are still higher than conventional aligners that can achieve an AER around 30 on the same dataset. There is still room for improvement in attention accuracy.

[2] The scores for E → C is much lower than C → E because BLEU is calculated at the word level rather than character level.

Word	Type	Freq.	Indep.	Joint
to	preposition	high	2.21	1.80
and	conjunction	high	2.21	1.60
the	definite article	high	1.96	1.56
yesterday	noun	medium	2.04	1.55
actively	adverb	medium	1.90	1.32
festival	noun	medium	1.55	0.85
inspects	verb	low	0.29	0.02
rebellious	adjective	low	0.29	0.02
noticing	verb	low	0.19	0.01

Table 4: Comparison of independent and joint training in terms of average attention entropy (see Eq. (15)) on Chinese-to-English translation.

4.5 Analysis of Alignment Matrices

We observe that a target word is prone to connect to too many source words in the alignment matrices produced by independent training. For example, in the lower alignment matrix of Figure 2(a), the third Chinese word "buxi" is aligned to three English words: "president", "bush", and "condemns". In addition, all the three alignment probabilities are relatively low. Similarly, four English words contribute to generating the last Chinese word "gongji": "condemns", "suicide", "boming", and "attack".

In contrast, agreement-based joint training leads to more concentrated alignment distributions. For example, in the lower alignment matrix of Figure 2(b), the third Chinese word "buxi" is most likely to be aligned to "bush". Likewise, the attention to the last Chinese word "gongji" now mainly focuses on "attack".

To measure the degree of concentration of attention, we define the *attention entropy* of a target word in a sentence pair as follows:

$$\mathrm{H}_{\mathbf{y}_n} = -\sum_{m=1}^{M} \mathbf{A}(\boldsymbol{\theta})_{n,m} \log \mathbf{A}(\boldsymbol{\theta})_{n,m} \quad (14)$$

Given a parallel corpus $D = \{\langle \mathbf{x}^{(s)}, \mathbf{y}^{(s)} \rangle\}_{s=1}^{S}$, the *aver-*

System	Training	Direction	Dev.	Test
MOSES	Indep.	E→F	28.38	32.31
		F→E	28.52	30.93
RNNSEARCH	Indep.	E→F	29.06	32.69
		F→E	28.32	29.99
	Joint	E→F	29.86^{**++}	33.45^{**++}
		F→E	29.01^{**++}	31.51^{**++}

Table 5: Results on the English-French translation task. The BLEU scores are case-insensitive. "**": significantly better than MOSES ($p < 0.01$); "++": significantly better than RNNSEARCH with independent training ($p < 0.01$).

age attention entropy is defined as

$$\tilde{\mathrm{H}}_y = \frac{1}{c(y,D)} \sum_{s=1}^{S} \sum_{n=1}^{N} \delta(\mathbf{y}_n^{(s)}, y) \mathrm{H}_{\mathbf{y}_n^{(s)}} \quad (15)$$

where $c(y, D)$ is the occurrence of a target word y on the training corpus D:

$$c(y,D) = \sum_{s=1}^{S} \sum_{n=1}^{N} \delta(\mathbf{y}_n^{(s)}, y) \quad (16)$$

Table 4 gives the average attention entropy of example words on the Chinese-to-English translation task. We find that the entropy generally goes downs with the decrease of word frequencies, which suggests that frequent target words tend to gain attention from multiple source words. Apparently, joint training leads to more concentrated attention than independent training. The gap seems to increase with the decrease of word frequencies.

4.6 Results on English-to-French Translation

Table 5 gives the results on the English-French translation task. While RNNSEARCH 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 NMT 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.

Luong et al. [2015a] 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 RNNSEARCH [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 RNNSEARCH. As our approach does not assume specific network architectures, it is possible to apply it to the models proposed by Luong et al. [2015a].

5.2 Agreement-based Learning

Liang et al. [2006] first introduce agreement-based learning into word alignment: encouraging 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 significantly improves both alignment and translation.

6 Conclusion

We have presented agreement-based joint training for bidirectional attention-based neural machine translation. By encouraging bidirectional 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.

Acknowledgements

This work was done while Yong Cheng and Shiqi Shen were visiting Baidu. This research is supported by the 973 Program (2014CB340501, 2014CB340505), the National Natural Science Foundation of China (No. 61522204, 61331013, 61361136003), 1000 Talent Plan grant, Tsinghua Initiative Research Program grants 20151080475 and a Google Faculty Research Award.

References

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

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

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

[Cho *et al.*, 2014] 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.

[Ganchev *et al.*, 2010] Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049, 2010.

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

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

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

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

[Koehn, 2004] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, 2004.

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

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

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

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

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

[Luong *et al.*, 2015a] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*, 2015.

[Luong *et al.*, 2015b] Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation. In *Proceedings of ACL*, 2015.

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

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

[Xu *et al.*, 2015] 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.