Incorporating Word Reordering Knowledge into Attention-based Neural Machine Translation

Jinchao Zhang[1] **Mingxuan Wang**[1] **Qun Liu**[3,1] **Jie Zhou**[2]
[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences
{zhangjinchao,wangmingxuan,liuqun}@ict.ac.cn
[2]Baidu Research - Institute of Deep Learning,
Baidu Inc.,Beijing,China
{zhoujie01}@baidu.com
[3]ADAPT Centre, School of Computing, Dublin City University

Abstract

This paper proposes three distortion models to explicitly incorporate the word reordering knowledge into attention-based Neural Machine Translation (NMT) for further improving translation performance. Our proposed models enable attention mechanism to attend to source words regarding both the semantic requirement and the word reordering penalty. Experiments on Chinese-English translation show that the approaches can improve word alignment quality and achieve significant translation improvements over a basic attention-based NMT by large margins. Compared with previous works on identical corpora, our system achieves the state-of-the-art performance on translation quality.

1 Introduction

Word reordering model is one of the most crucial sub-components in Statistical Machine Translation (SMT) (Brown et al., 1993; Koehn et al., 2003; Chiang, 2005) which provides word reordering knowledge to ensure reasonable translation order of source words. It is separately trained and then incorporated into the SMT framework in a pipeline style.

In recent years, end-to-end NMT (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015) has made tremendous progress (Jean et al., 2015; Luong et al., 2015b; Shen et al., 2016; Sennrich et al., 2016; Tu et al., 2016; Zhou et al., 2016; Johnson et al., 2016). An encoder-decoder framework (Cho et al., 2014b; Sutskever et al., 2014) with attention mechanism (Bahdanau et al., 2015) is widely used, in which an encoder compresses the source sentence, an attention mechanism evaluates related source words and a decoder generates target words.

The attention mechanism evaluates the distribution of to-be-translated source words in a content-based addressing fashion (Graves et al., 2014) which tends to attend to the source words regarding the content relation with current translation status. Lack of explicit models to exploit the word reordering knowledge may lead to attention faults and generate fluent but inaccurate or inadequate translations. Table 1 shows a translation instance and Figure 1 depicts the corresponding word alignment matrix that produced by the attention mechanism. In this example, even though the word "zuixin (latest)" is a common adjective in Chinese and its following word should be translated soon in Chinese to English translation direction, the word "yiju (evidence)" does not obtain appropriate attention which leads to the incorrect translation.

src	youguan(related) baodao(report) shi(is) zhichi(support) tamen(their) lundian(arguments) de('s) zuixin(latest) yiju(evidence) .
ref	the report is the latest evidence that supports their arguments .
NMT	the report supports their perception of the latest .
count	zuixin yiju {0}

Table 1: An instance in Chinese-English translation task. The row "count" represents the frequency of the word collocation in the training corpus. The collocation "zuixin yiju" does not appear in the training data.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1524–1534
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1140

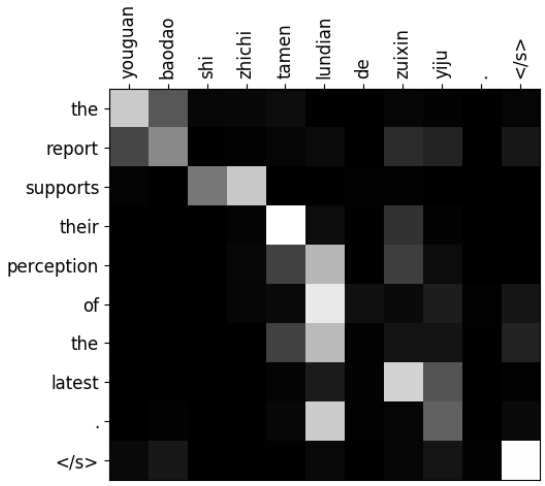


Figure 1: The source word "yiju" does not obtain appropriate attention and its word sense is completely neglected.

To enhance the attention mechanism, implicit word reordering knowledge needs to be incorporated into attention-based NMT. In this paper, we introduce three distortion models that originated from SMT (Brown et al., 1993; Koehn et al., 2003; Och et al., 2004; Tillmann, 2004; Al-Onaizan and Papineni, 2006), so as to model the word reordering knowledge as the probability distribution of the relative jump distances between the newly translated source word and the to-be-translated source word. Our focus is to extend the attention mechanism to attend to source words regarding both the semantic requirement and the word reordering penalty.

Our models have three merits:

1. *Extended word reordering knowledge.* Our models capture explicit word reordering knowledge to guide the attending process for attention mechanism.

2. *Convenient to be incorporated into attention-based NMT.* Our distortion models are differentiable and can be trained in the end-to-end style. The interpolation approach ensures that the proposed models can coordinately work with the original attention mechanism.

3. *Flexible to utilize variant context for computing the word reordering penalty.* In this paper, we exploit three categories of information as distortion context conditions to compute the word reordering penalty, but variant context information can be utilized due to our model's flexibility.

We validate our models on the Chinese-English translation task and achieve notable improvements:

- On 16K vocabularies, NMT models are usually inferior in comparison with the phrase-based SMT, but our model surpasses phrase-based Moses by average **4.43** BLEU points and outperforms the attention-based NMT baseline system by **5.09** BLEU points.

- On 30K vocabularies, the improvements over the phrase-based Moses and the attention-based NMT baseline system are average **6.06** and **1.57** BLEU points respectively.

- Compared with previous work on identical corpora, we achieve the state-of-the-art translation performance on average.

The word alignment quality evaluation shows that our model can effectively improve the word alignment quality that is crucial for improving translation quality.

2 Background

We aim to capture word reordering knowledge for the attention-based NMT by incorporating distortion models. This section briefly introduces attention-based NMT and distortion models in SMT.

2.1 Attention-based Neural Machine Translation

Formally, given a source sentence $\mathbf{x} = \mathbf{x}_1, ..., \mathbf{x}_m$ and a target sentence $\mathbf{y} = \mathbf{y}_1, ..., \mathbf{y}_n$, NMT models the translation probability as

$$P(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{n} P(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{x}), \quad (1)$$

where $\mathbf{y}_{<t} = \mathbf{y}_1, ..., \mathbf{y}_{t-1}$. The generation probability of $\mathbf{y}_t$ is

$$P(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{x}) = g(\mathbf{y}_{t-1}, \mathbf{c}_t, \mathbf{s}_t), \quad (2)$$

where $g(\cdot)$ is a softmax regression function, $\mathbf{y}_{t-1}$ is the newly translated target word and

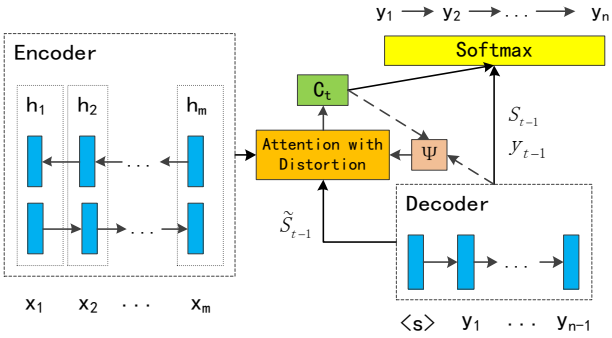


Figure 2: The general architecture of our proposed models. The dash line represents variant context can be utilized to determine the word reordering penalty.

$\mathbf{s}_t$ is the hidden states of decoder which represents the translation status.

The attention $\mathbf{c}_t$ denotes the related source words for generating $\mathbf{y}_t$ and is computed as the weighted-sum of source representation $\mathbf{h}$ upon an alignment vector α_t shown in Eq.(3) where the $align(\cdot)$ function is a feedforward network with $softmax$ normalization.

$$\begin{aligned} \mathbf{c}_t &= \sum_{j=1}^{m} \alpha_{t,j}\mathbf{h}_j \\ \alpha_{t,j} &= align(\mathbf{s}_t, \mathbf{h}_j) \end{aligned} \tag{3}$$

The hidden states $\mathbf{s}_t$ is updated as

$$\mathbf{s}_t = f(\mathbf{s}_{t-1}, \mathbf{y}_{t-1}, \mathbf{c}_t), \tag{4}$$

where $f(\cdot)$ is a recurrent function.

We adopt a varietal attention mechanism[1] in our in-house RNNsearch model which is implemented as

$$\begin{aligned} \widetilde{\mathbf{s}}_t &= f_1(\mathbf{s}_{t-1}, \mathbf{y}_{t-1}), \\ \alpha_{t,j} &= align(\widetilde{\mathbf{s}}_t, \mathbf{h}_j), \\ \mathbf{s}_t &= f_2(\widetilde{\mathbf{s}}_t, \mathbf{c}_t), \end{aligned} \tag{5}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are recurrent functions.

As shown in Eq.(3), the attention mechanism attends to source words in a content-based addressing way without considering any explicit word reordering knowledge. We introduce distortion models to capture explicit word reordering knowledge for enhancing the attention mechanism and improving translation quality.

[1]https://github.com/nyu-dl/dl4mt-tutorial/tree/master/session2

2.2 Distortion Models in SMT

In SMT, distortion models are linearly combined with other features, as follows,

$$\begin{aligned} \mathbf{y}^* = \arg\max_{\mathbf{y}} \exp[&\lambda_d d(\mathbf{x}, \mathbf{y}, b) + \\ &\sum_{r=1}^{R-1} \lambda_r h_r(\mathbf{x}, \mathbf{y}, b)], \end{aligned} \tag{6}$$

where $d(\cdot)$ is the distortion feature, $h_r(\cdot)$ represents other features, λ_d and λ_r are the weights, b is the latent variable that represents translation knowledge and R is the number of features.

IBM Models (Brown et al., 1993) depicted the word reordering knowledge as positional relations between source and target words. Koehn et al. (2003) proposed a distortion model for phrase-based SMT based on jump distances between the newly translated phrases and to-be-translated phrases which does not consider specific lexical information. Och et al. (2004) and Tillmann (2004) proposed orientation-based distortion models that consider translation orientations. Yaser and Papineni (2006) proposed a distortion model to estimate probability distribution on possible relative jumps conditioned on source words.

These models are proposed for SMT and separately trained as sub-components. Inspired by these previous work, we introduce the distortion models into NMT model for modeling the word reordering knowledge. Our proposed models are designed for NMT which can be trained in the end-to-end style.

3 Distortion Models for attention-based NMT

The basic idea of our proposed distortion models is to estimate the probability distribution of the possible relative jump distances between the newly translated source word and the to-be-translated source word upon the context condition. Figure 2 shows the general architecture of our proposed model.

3.1 General Architecture

We employ an interpolation approach to incorporate distortion models into attention-based NMT as

$$\alpha_t = \lambda \cdot \mathbf{d}_t + (1-\lambda)\hat{\alpha}_t, \tag{7}$$

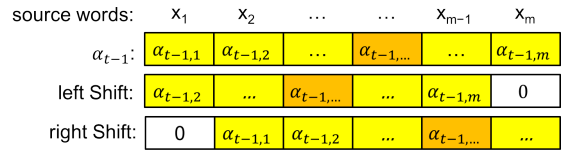


Figure 3: Illustration of shift actions of the alignment vector α_{t-1}. If α_t is the left shift of α_{t-1}, it represents the translation orientation of the source sentence is backward and if α_t is the right shift of α_{t-1}, the translation orientation is forward.

where α_t is the ultimate alignment vector for computing the related source context $\mathbf{c}_t$, $\mathbf{d}_t$ is the alignment vector calculated by the distortion model, $\hat{\alpha}_t$ is the alignment vector computed by the basic attention mechanism and λ is a hyper-parameter to control the weight of the distortion model.

In the proposed distortion model, relative jumps on source words are depicted as the "shift" actions of the alignment vector α_{t-1} which is shown in the Figure 3. The right shift of α_{t-1} indicates that the translation orientation of source words is forward and the left shift represents that the translation orientation is backward. The extent of a shift action measures the word reordering distance. Alignment vector $\mathbf{d}_t$, which is produced by the distortion model, is the expectation of all possible shifts of α_{t-1} conditioned on certain context.

Formally, the proposed distortion model is

$$\begin{aligned}\mathbf{d}_t &= E[\Gamma(\alpha_{t-1})] \\ &= \sum_{k=-l}^{l} P(k|\Psi) \cdot \Gamma(\alpha_{t-1}, k),\end{aligned} \tag{8}$$

where $k \in [-l, l]$ is the possible relative jump distance, l is the window size parameter and $P(k|\Psi)$ stands for the probability of jump distance k that conditioned on the context Ψ. Function $\Gamma(\cdot)$ for shifting the alignment vector is defined as

$$\Gamma(\alpha_{t-1}, k) = \begin{cases} \{\alpha_{t-1,-k}, ..., \alpha_{t-1,m}, 0, ..., 0\}, & \text{k<0} \\ \alpha_{t-1}, & \text{k= 0} \\ \{0, ..., 0, \alpha_{t-1,1}, ..., \alpha_{t-1,m-k}\}, & \text{k>0} \end{cases} \tag{9}$$

which can be implemented as matrix multiplication computations.

We respectively exploit source context, target context and translation status context (hidden states of decoder) as Ψ and derive three distortion models: Source-based Distortion (**S-Distortion**) model , Target-based Distortion (**T-Distortion**) model and Translation-status-based Distortion (**H-Distortion**) model. Our framework is capable of utilizing arbitrary context as the condition Ψ to predict the relative jump distances.

3.2 S-Distortion model

S-Distortion model adopts previous source context $\mathbf{c}_{t-1}$ as the context Ψ with the intuition that certain source word indicate certain jump distance. The to-be-translated source word have intense positional relations with the newly translated one.

The underlying linguistic intuition is that synchronous grammars (Yamada and Knight, 2001; Galley et al., 2004) can be extracted from language pairs. Word categories such as verb, adjective and preposition carry general word reordering knowledge and words carry specific word reordering knowledge.

To further illustrate this idea, we present some common synchronous grammar rules that can be extracted from the example in Table 1 as follows,

$$\begin{aligned} &NP \longrightarrow JJ \quad NN \,|\, JJ \quad NN \\ &JJ \longrightarrow zuixin \,|\, latest. \end{aligned} \tag{10}$$

From the above grammar, we can conjecture the speculation that after the word "zuixin(latest)" is translated, the translation orientation is forward with shift distance 1.

The probability function in S-Distortion model is defined as follows,

$$\begin{aligned} P(\cdot|\Psi) &= z(\mathbf{c}_{t-1}) \\ &= softmax(W_c \mathbf{c}_{t-1} + b_c), \end{aligned} \tag{11}$$

where $W_c \in \mathbb{R}^{(2l+1)\times dim(c_{t-1})}$ and $b_c \in \mathbb{R}^{2l+1}$ are weight matrix and bias parameters.

3.3 T-Distortion Model

T-Distortion model exploits the embedding of the previous generated target word $\mathbf{y}_{t-1}$ as the context condition to predict the probability distribution of distortion distances. It focuses on the word reordering knowledge upon target

word context. As illustrated in Eq.(10), the target word "latest" possesses word reordering knowledge that is identical with source word "zuixin".

The probability function in T-Distortion model is defined as follows,

$$\begin{aligned} P(\cdot|\Psi) &= z(\mathbf{y}_{t-1}) \\ &= softmax(W_y emb(\mathbf{y}_{t-1}) + b_{\mathbf{y}}), \end{aligned} \tag{12}$$

where $emb(\mathbf{y}_{t-1})$ is the embedding of $\mathbf{y}_{t-1}$, $W_y \in \mathbb{R}^{(2l+1)\times dim(emb(\mathbf{y}_{t-1}))}$ and $b_{\mathbf{y}} \in \mathbb{R}^{2l+1}$ are weight matrix and bias parameters.

3.4 H-Distortion Model

The hidden states $\tilde{s}_{t-1}$ reflect the translation status and contains both source context and target context information. Therefore, we exploit $\tilde{s}_{t-1}$ as context Ψ in the H-Distortion model to predict shift distances.

The probability function in H-Distortion model is defined as follows,

$$\begin{aligned} P(\cdot|\Psi) &= z(\tilde{s}_{t-1}) \\ &= softmax(W_s \tilde{s}_{t-1} + b_s) \end{aligned} \tag{13}$$

where $W_s \in \mathbb{R}^{(2l+1)\times dim(\tilde{s}_{t-1})}$ and $b_s \in \mathbb{R}^{2l+1}$ are the weight matrix and bias parameters.

4 Experiments

We carry the translation task on the Chinese-English direction to evaluate the effectiveness of our models. To investigate the word alignment quality, we take the word alignment quality evaluation on the manually aligned corpus. We also conduct the experiments to observe effects of hyper-parameters and the training strategies.

4.1 Data and Metrics

Data: Our Chinese-English training corpus consists of 1.25M sentence pairs extracted from LDC corpora[2] with 27.9M Chinese words and 34.5M English words respectively. 16K vocabularies cover approximately 95.8% and 98.3% words and 30K vocabularies cover approximately 97.7% and 99.3% words in Chinese and English respectively. We choose NIST 2002 dataset as the validation set. NIST 2003-2006 are used as test sets. To assess the word alignment quality, we employ Tsinghua dataset (Liu and Sun, 2015) which contains 900 manually aligned sentence pairs.

Metrics: The translation quality evaluation metric is the case-insensitive 4-gram BLEU[3] (Papineni et al., 2002). *Sign-test* (Collins et al., 2005) is exploited for statistical significance test. Alignment error rate (AER) (Och and Ney, 2003) is calculated to assess the word alignment quality.

4.2 Comparison Systems

We compare our approaches with three baseline systems:

Moses (Koehn et al., 2007): An open source phrase-based SMT system with default settings. Words are aligned with GIZA++ (Och and Ney, 2003). The 4-gram language model with modified Kneser-Ney smoothing is trained on the target portion of training data by SRILM (Stolcke et al., 2002).

Groundhog[4]: An open source attention-based NMT system with default settings.

RNNsearch*: Our in-house implementation of NMT system with the varietal attention mechanism and other settings that presented in section 4.3.

4.3 Training

Hyper parameters: The sentence length for training NMTs is up to 50, while SMT model exploits whole training data without any restrictions. Following Bahdanau et al. (2015), we use bi-directional Gated Recurrent Unit (GRU) as the encoder. The forward representation and the backward representation are concatenated at the corresponding position as the ultimate representation of a source word. The word embedding dimension is set to 620 and the hidden layer size is 1000. The interpolation parameter λ is 0.5 and the window size l is set to 3.

Training details:

Square matrices are initialized in a random orthogonal way. Non-square matrices are initialized by sampling each element from the

[2]The corpora includes LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06.

[3]ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl

[4]https://github.com/lisa-groundhog/GroundHog

Systems	MT03	MT04	MT05	MT06	Average	Average Increase
Moses	31.61	33.48	30.75	30.85	31.67	–
Groundhog(16K)	29.14	31.23	28.11	27.77	29.06	–
RNNsearch*(16K)	30.77	33.92	30.82	28.56	31.02	–
+ T-Distortion	35.71‡	37.81‡	33.78‡	33.79‡	35.27	+(**4.26**, **3.60**, **6.21**)
+ S-Distortion	**36.58**‡	38.47 ‡	34.85‡	33.86‡	35.94	+(**4.92**, **4.27**, **6.88**)
+ H-Distortion	35.95‡	**38.77**‡	**35.33**‡	**34.36**‡	**36.10**	+(**5.09**, **4.43**, **7.04**)
Groundhog(30K)	31.92	34.09	31.56	31.12	32.17	–
RNNsearch*(30K)	36.47	39.17	35.04	33.97	36.16	–
+ T-Distortion	37.93†	40.40‡	**36.81**‡	**35.77**‡	**37.73**	+(**1.57**, **6.06**, **5.56**)
+ S-Distortion	37.47†	**40.52**‡	36.16‡	35.32	37.37	+(**1.21**, **5.70**, **5.20**)
+ H-Distortion	**38.33**‡	40.11‡	36.71†	35.29‡	37.61	+(**1.45**, **5.94**, **5.44**)

Table 2: BLEU-4 scores (%) on NIST test set 03-06 of Moses (default settings), Groundhog (default settings), RNNsearch* and RNNsearch* with distortion models respectively. The values in brackets are increases on RNNsearch*, Moses and Groundhog respectively. ‡ indicates statistical significant difference ($p<0.01$) from RNNsearch* and † means statistical significant difference ($p<0.05$) from RNNsearch*.

Gaussian distribution with mean 0 and variance 0.01^2. All bias are initialized to 0.

Parameters are updated by Mini-batch Gradient Descent and the learning rate is controlled by the AdaDelta (Zeiler, 2012) algorithm with decay constant $\rho = 0.95$ and denominator constant $\epsilon = 1e-6$. The batch size is 80. Dropout strategy (Srivastava et al., 2014) is applied to the output layer with the dropout rate 0.5 to avoid over-fitting. The gradients of the cost function which have $L2$ norm larger than a predefined threshold 1.0 is normalized to the threshold to avoid gradients explosion (Pascanu et al., 2013). We exploit length normalization (Cho et al., 2014a) on candidate translations and the beam size for decoding is 12. For NMT with distortion models, we use trained RNNsearch* model to initialize parameters except for those related to distortions.

4.4 Results

The translation quality experiment results are shown in Table 2. We carry the experiments on different vocabulary sizes for that different vocabulary sizes cause different degrees of the rare word collocations. Through this way, we can validate the effects of our proposed models in alleviating the rare word collocations problem that leads to incorrect word alignments.

On 16K vocabularies: The phrase-based Moses performs better than the basic NMTs including Groundhog and RNNsearch*. Besides the differences between model architectures, restricted vocabularies and sentence length also affect the performance of NMTs. However, RNNsearch* with distortion models surpass phrase-based Moses by average 3.60, 4.27 and 4.43 BLEU points. RNNsearch* outperforms Groundhog by average 1.96 BLEU points due to the varietal attention mechanism, length normalization and dropout strategies. Distortion models bring about remarkable improvements as 4.26, 4.92 and 5.09 BLEU points over the RNNsearch* model.

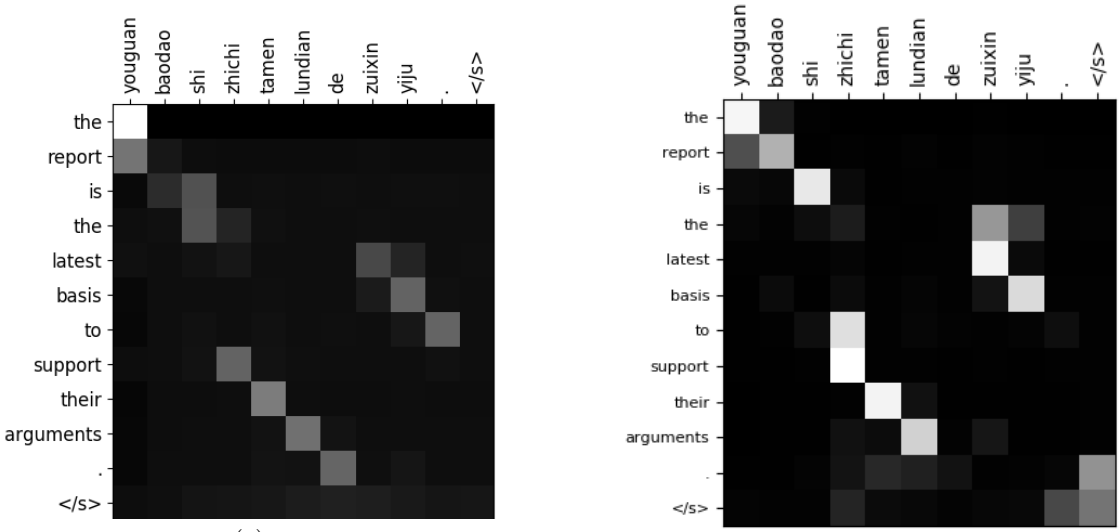
On 30K vocabularies: RNNsearch* with distortion models yield average gains by 1.57, 1.21 and 1.45 BLEU points over RNNsearch* and outperform phrase-based Moses by average 6.06, 5.70 and 5.94 BLEU points and surpass GroundHog by average 5.56, 5.20 and 5.44 BLEU points. RNNsearch*(16K) with distortion models achieve close performances with RNNsearch*(30K). The improvements on 16K vocabularies are larger than that on 30K vocabularies for the intuition that more "UNK" words lead to more rare word collocations, which results in serious attention ambiguities.

The RNNsearch* with distortion models yield tremendous improvements on BLEU scores proves the effectiveness of proposed approaches in improving translation quality.

Comparison with previous work: We present the performance comparison with pre-

System	Length	MT03	MT04	MT05	MT06	Average
Coverage	80	-	-	32.73	32.47	-
MEMDEC	50	36.16	39.81	35.91	**35.98**	36.95
NMT_{IA}	80	35.69	39.24	35.74	35.10	36.44
Our work	50	**37.93**	**40.40**	**36.81**	35.77	**37.73**

Table 3: Comparison with previous work on identical training corpora. Coverage (Tu et al., 2016) is a basic RNNsearch model with a coverage model to alleviate the over-translation and under-translation problems. MEMDEC (Wang et al., 2016) is to improve translation quality with external memory. NMT_{IA} (Meng et al., 2016) exploits a readable and writable attention mechanism to keep track of interactive history in decoding. Our work is NMT with H-Distortion model. The vocabulary sizes of all work are 30K and maximum lengths of sentence differ.



(a) (b)

Figure 4: (a) is the output of the distortion model and is calculated on shift actions of previous alignment vector. (b) is the ultimate word alignment matrix of attention-based NMT with H-Distortion model. Compared with Figure 1, (b) is more centralized and accurate.

Systems	BLEU	AER
RNNsearch*(30K)	20.90	49.73
+ T-Distortion	24.33‡	46.92
+ S-Distortion	24.10‡	47.37
+ H-Distortion	24.42‡	47.05

Table 4: BLEU-4 scores (%) and AER scores on Tsinghua manually aligned Chinese-English evaluation set. The lower the AER score, the better the alignment quality.

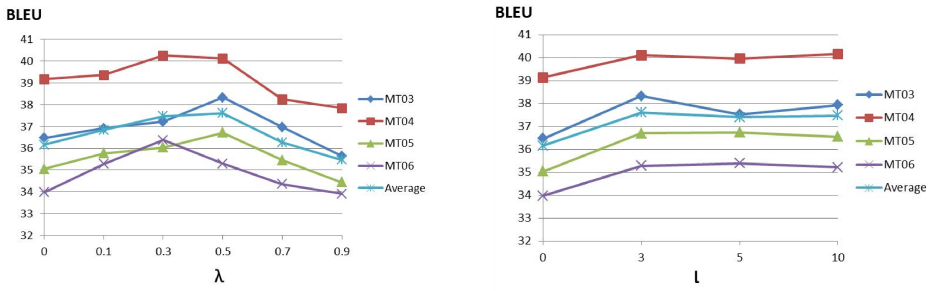
vious work that employ identical training corpora in Table 3. Our work evidently outperforms previous work on average performance. Although we restrict the maximum length of sentence to 50, our model achieves the state-of-the-art BLEU scores on almost all test sets except NIST2006.

4.5 Analysis

We investigate the effects on the alignment quality of our models and conduct the experiments to evaluate the influence of the hyper-parameter settings and the training strategies.

4.5.1 Alignment Quality

Distortion models concentrate on attending to to-be-translated words based on the word reordering knowledge and can intuitively enhance the word alignment quality. To investigate the effect on word alignment quality, we apply the BLEU and AER evaluations on Tsinghua manually aligned data set.



(a) (b)

Figure 5: Translation performance on the test sets with respect to the hyper-parameter λ and l.

System	MT03	MT04	MT05	MT06	Average
Pre-training	35.95	**38.77**	**35.33**	**34.36**	**36.10**
No pre-training	**36.99**	38.42	34.56	34.01	36.00

Table 5: Comparison between pre-training and no pre-training H-Distortion model. The performances are consistent.

Table 4 lists the BLEU and AER scores of Chinese-English translation with 30K vocabulary. RNNsearch*(30K) with distortion models achieve significant improvements on BLEU scores and obvious decrease on AER scores. The results shows that the proposed model can effectively improve the word alignment quality

Figure 4 shows the output of distortion model and ultimate alignment matrix of the above-mentioned instance. Compared with Figure 1, the alignment matrix produced by NMT with distortion models is more concentrated and accurate. The output of distortion model shows its capacity of modeling word reordering knowledge.

4.5.2 Effect of Hyper-parameters

To investigate the effect of the weight hyper-parameter λ and window hyper-parameter l in the proposed model, we carry experiments on H-Distortion model with variable hyper-parameter settings. We fix $l = 3$ for exploring the effect of λ and fix $\lambda = 0.5$ for observing the effect of l. Figure 5 presents the translation performances with respect to hyper-parameters. With the increase of weight λ, the BLEU scores first rise and then drop, which shows the distortion model provides additional helpful information while can not fully cover the attention mechanism for its insufficient content searching ability. For window l, the experiments show that larger windows bring slight further improvements, which indicates that distortion model pays more attention to the short-distance reordering knowledge.

4.5.3 Pre-training VS No Pre-training

We conduct the experiment without using pre-training strategy to observe the effect of the initialization. As is shown in Table 5, the no-pre-training model achieves consistent improvements with the pre-training one which verifies the stable effectiveness of our approach. Initialization with pre-training strategy provides a fast approach to obtain the model for it needs fewer training iterations.

5 Related Work

Our work is inspired by the distortion models that widely used in SMT. The most related work in SMT is the distortion model proposed by Yaser and Papineni (2006). Their model is identical to our S-Distortion model that captures the relative jump distance knowledge on source words. However, our approach is deliberately designed for the attention-based NMT system and is capable of exploiting variant context information to predict the relative jump distances.

Our work is related to the work (Luong et al., 2015a; Feng et al., 2016; Tu et al., 2016;

Cohn et al., 2016; Meng et al., 2016; Wang et al., 2016) that concentrate on the improvement of the attention mechanism. To remit the computing cost of the attention mechanism when dealing with long sentences, Luong et al. (2015a) proposed the local attention mechanism by just focusing on a subscope of source positions. Cohn et al. (2016) incorporated structural alignment biases into the attention mechanism and obtained improvements across several challenging language pairs in low-resource settings. Feng et al. (2016) passed the previous attention context to the attention mechanism by adding recurrent connections as the implicit distortion model. Tu et al. (2016) maintained a coverage vector for keeping the attention history to acquire accurate translations. Meng et al. (2016) proposed the interactive attention with the attentive read and attentive write operation to keep track of the interaction history. Wang et al. (2016) utilized an external memory to store additional information for guiding the attention computation. These works are different from ours, as our distortion models explicitly capture word reordering knowledge through estimating the probability distribution of relative jump distances on source words to incorporate word reordering knowledge into the attention-based NMT.

6 Conclusions

We have presented three distortion models to enhance attention-based NMT through incorporating the word reordering knowledge. The basic idea of proposed distortion models is to enable the attention mechanism to attend to the source words regarding both semantic requirement and the word reordering penalty. Experiments show that our models can evidently improve the word alignment quality and translation performance. Compared with previous work on identical corpora, our model achieves the state-of-the-art performance on average. Our model is convenient to be applied in the attention-based NMT and can be trained in the end-to-end style. We also investigated the effect of hyper-parameters and pre-training strategy and further proved the stable effectiveness of our model. In the future, we plan to validate the effectiveness of our model on more language pairs.

7 Acknowledgement

Qun Liu's work is partially supported by Science Foundation Ireland in the ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Dublin City University funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund. We are grateful to Qiuye Zhao, Fandong Meng and Daqi Zheng for their helpful suggestions. We thank the anonymous reviewers for their insightful comments.

References

Yaser Al-Onaizan and Kishore Papineni. 2006. Distortion models for statistical machine translation. In *Proceedings of ACL2006*. pages 529–536.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR2015*.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation:parameter estimation. *Computational Linguistics* 19(2):263—311.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL2005*. pages 263–270.

Kyunghyun Cho, Bart Van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: Encoder–decoder approaches. In *Eighth Workshop on Syntax,Semantics and Structure in Statistical Translation*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of EMNLP 2014*. Doha, Qatar, pages 1724–1734.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of NAACL2016*. pages 876—885.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL2005*. pages 531–540.

Shi Feng, Shu jie Liu, Mu Li, and Ming Zhou. 2016. Implicit distortion and fertility models for attention-based encoder-decoder nmt model. *arXiv preprint arXiv:1601.03317* .

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule. In *Proceedings of HLT/NAACL*. Boston, volume 4, pages 273–280.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401* .

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of ACL2014*. volume 1, pages 1–10.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, and Greg Corrado. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. In *arXiv preprint arXiv:1609.08144*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of EMNLP2013*. Seattle, Washington, USA, pages 1700–1709.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL2007 Demo and Poster Sessions*. Prague, Czech Republic, pages 177–180.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings NAACL2003*. pages 48–54.

Yang Liu and Maosong Sun. 2015. Contrastive unsupervised word alignment with non-local features. In *Proceedings of AAAI2015*. pages 2295–2301.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015a. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP2015*. Lisbon, Portugal.

Minh Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2015b. Addressing the rare word problem in neural machine translation. *Proceedings of ACL2015* 27(2):82–86.

Fandong Meng, Zhengdong Lu, Hang Li, and Qun Liu. 2016. Interactive attention for neural machine translation. In *Proceedings of COLING2016*.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alexander M Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, et al. 2004. A smorgasbord of features for statistical machine translation. In *HLT-NAACL*. pages 161–168.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics* 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL2002*. Association for Computational Linguistics, pages 311–318.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. *ICML (3)* 28:1310–1318.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL2016*. pages 1715–1725.

Shiqi Shen, Yong Cheng, Zhongjun He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of ACL2016*. pages 1683–1692.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing*. volume 2, pages 901–904.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS2014*.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*. pages 101–104.

Zhaopeng Tu, Zhengdong Lu, yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of ACL*. pages 76–85.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2016. Memory-enhanced decoder for neural machine translation. In *Proceedings of EMNLP2016*.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of ACL2001*. pages 523–530.

Al-Onaizan Yaser and Kishore Papineni. 2006. Distortion models for statistical machine translation. In *Proceedings of ACL2006*. pages 529–536.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* .

Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. 2016. Deep recurrent models with fast-forward connections for neural machine translation. In *Proceedings of EMNLP2016*.