Information-Propagation-Enhanced Neural Machine Translation by Relation Model

Wen Zhang\(^1\)  Jiawei Hu\(^1\)  Yang Feng\(^1\)  Qun Liu\(^1,2\)

\(^1\)Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS
\{zhangwen,hujiawei,fengyang\}@ict.ac.cn
\(^2\)ADAPT Centre, School of Computing, Dublin City University
qliu@computing.dcu.ie

Abstract
Even though sequence-to-sequence neural machine translation (NMT) model have achieved state-of-art performance in the recent fewer years, but it is widely concerned that the recurrent neural network (RNN) units are very hard to capture the long-distance state information, which means RNN can hardly find the feature with long term dependency as the sequence becomes longer. Similarly, convolutional neural network (CNN) is introduced into NMT for speeding recently, however, CNN focus on capturing the local feature of the sequence; To relieve this issue, we incorporate a relation network into the standard encoder-decoder framework to enhance information-propogation in neural network, ensuring that the information of the source sentence can flow into the decoder adequately. Experiments show that proposed framework outperforms the statistical MT model and the state-of-art NMT model significantly on two data sets with different scales.

Introduction
In recent years, neural machine translation (NMT) (Cho et al. 2014b; Sutskever, Vinyals, and Le 2014; Cho et al. 2014a) has achieved great success in some language pairs, over the state-of-the-art statistical machine translation (SMT) (Koehn, Hoang, and Birch 2007). The encoder-decoder architecture is widely used for NMT, the principle behind which is that: encoding the meaning of the input into a concept space and performing translation based on this encoding. This ‘meaning encoding’ principle leads to a deeper understanding and learning of the translation rules, and hence a better translation than conventional statistic machine translation (SMT) that considers only surface forms, i.e., words and phrases.

However, even attention-based NMT uses bi-directional RNN to encode the source sentence in two directions, the representation contains only the sequential relationship among the words in the source sentence, and, as the length of the source sentence increases, the encoded vector may forget the information in the words far away from it, bi-directional encoder may solve this problem in some degree; However, sentences in natural language are regarded as relational, which means it may have grammatical relation (such as dependency relationship) between some word and another in one sentence, recurrent neural network models like GRU or LSTM are very difficult to capture these kinds of relations between word-pairs in one sentence, thus, when making dynamic alignment, the attention model is also hard to recognize this relation; To address this problem, (Bastings et al. 2017) introduces graph-convolutinal networks (GCNs) into the encoder of attention-based NMT, which simultaneously take the representation and the syntactic dependency tree of the source sentence as input to produce another representation for each word, beneficially, these representations may be sensitive to their syntactic neighborhoods.

In this paper, we propose to introduce another simple neural network called Relation Networks (RNs) (Santoro et al. 2017) into the attention module of attention-based NMT compatibly, to let the NMT model capture the relations between each word-pair in the source sentence during the process of dynamic alignment. We validate our proposed model on NIST Chinese-English translation task. Experiment results show that, after incorporated with RN, the attention-based NMT model can generate a entangled sequential representation of source sentence by recognizing relations between words, then build higher quality dynamic alignment according to these relations, finally, produce more accurate performance for machine translation further.

Background
Based on RNNSearch, we build our machine translation system; RNNSearch improved the attention mechanism of attention-based NMT (Bahdanau, Cho, and Bengio 2014), which first represents the source sentence into a sequence of vectors by an bi-directional RNN encoder, then another RNN decoder learns to align and generates target translation word by word, in the process of which, a feed-forward neural network is applied to produce dynamic alignment according to the representation of the source sentence, previous target word and the previous hidden state of the decoder. We start this section by describing the attention-based NMT model.

Attention-based Neural Machine Translation
Figure 1 shows the framework of attention-based NMT, which is composed of three components: Encoder, Attention layer and Decoder.
Generally, given a source sequence \( x = \{x_1, \cdots, x_{l_x}\} \) and the incomplete target sequence \( y_{<j} = \{y_1, \cdots, y_{j-1}\} \) which have been generated. The model predict next word according following target-word probability distribution:

\[
p(y_j|x,y_{<j}) \propto \exp(f(y_{j-1}, s_j, a_j) \times W_v) \tag{1}
\]

where function \( f \) is a recurrent neural network, \( W_v \) is a mapping weighted matrix to map the output of \( f \) into the dimension of target vocabulary size, which is limited to the most frequent 15K words (Cho et al. 2014b). Then, we apply softmax function to the final product to get the probability distribution over the whole target vocabulary for each target word.

\( a_j \) can be treated as the context, which is actually the weighted sum of each annotation \( h_i \) in the source sentence, computed by the attention model:

\[
a_j = \sum_{i=1}^{l_x} \alpha_{ij} h_i \tag{2}
\]

Bi-directional RNN encoder embodies two RNNs (here we use GRUs), which respectively encode \( x \) into a sequence of fixed-length vectors \( \overrightarrow{h} = \{\overrightarrow{h}_1, \cdots, \overrightarrow{h}_{l_x}\} \) from left to right and \( \overleftarrow{h} = \{\overleftarrow{h}_1, \cdots, \overleftarrow{h}_{l_x}\} \) from right to left, each \( h_i \) is the concatenation of \( \overrightarrow{h}_i \) and \( \overleftarrow{h}_i \):

\[
h_i = \begin{bmatrix} \overrightarrow{h}_i \\ \overleftarrow{h}_i \end{bmatrix} = \begin{bmatrix} \text{GRU} \left( x_i, \overrightarrow{h}_{i-1} \right) \\ \text{GRU} \left( x_i, \overleftarrow{h}_{i+1} \right) \end{bmatrix} \tag{3}
\]

\( h_i \) contains the information about \( x_i \) and all other words in the source sentence with a strong focus on the parts surrounding \( x_i \).

\( \alpha_{ij} \) is the weight of \( h_i \), and can be considered as probability that the \((j-1)\)th word in the target sentence is aligned to the \( j \)th word in the source sentence, it is computed by

\[
\alpha_{ij} = \frac{\exp[g(s_{j-1}, h_i)]}{\sum_{i=1}^{l_x} \exp[g(s_{j-1}, h_i)]} \tag{4}
\]

where we follow the definition of function \( g \) in (Bahdanau, Cho, and Bengio 2014)

\[
g(s_{j-1}, h_i) = v^T_j \tanh(W_a s_{j-1} + U_a h_i) \tag{5}
\]

g is actually a feed-forward network, which can measure how well previous hidden state \( s_{j-1} \) and \( h_i \) match with each other, hidden state \( s_j \) is calculated by

\[
s_j = \text{GRU} (y_{j-1}, s_{j-1}, a_j) \tag{6}
\]

**Improved Decoder**

The improved decoder model is shown in Figure 2, we only replace the decoder part in Figure 1 with the improved counterpart, leaving other modules unaltered. In the new decoder, a new intermediate hidden state \( s'_j \) is introduced, and the final hidden state \( s_j \) is computed based on \( s'_j \); So \( s_j \) would not be calculated by \( y_{j-1}, s_{j-1} \) and \( a_j \) directly, Formula (6) is decomposed into two others:

One GRU computes the intermediate hidden state \( s'_j \) according to \( y_{j-1} \) and \( s_{j-1} \):

\[
s'_j = \text{GRU}'(y_{j-1}, s_{j-1}) \tag{7}
\]

Then, the other one is used to calculate next final hidden state based on the intermediate state and attention

\[
s_j = \text{GRU}(s'_j, a_j) \tag{8}
\]

Differently, Formula (4) and (5) are also updated, instead of \( s_{j-1}, s'_j \) is used to generate \( a_{ij} \) together with \( h_i \), as shown by Attention Layer in Figure 2.

**Relational Attention Model**

Based on improved attention model in Figure 2, we incorporate another relation layer between the attention layer and bi-directional encoder of NMT; As shown in Figure 3, the relation layer is composed of three components:

- Convolutional Encoder (CENC)
- Graph Propogation Layer (GPL)
- Multi-Layer Perceptron Decoder (MLPDEC)

**CENC:** the initial representation of the source sequence generated by NMT encoder are convolved through two layers 1-dimensional convolutional neural network (CNN) on the dimension of sentence length, with the length unchanged, so, after that, the encoder outputs \( d \times k \)-dimensional feature maps, which is a new convolutional representation for each source word.

**GPL:** The graph propogation layer concatenates the convolutional representations of each source word-pair, it is this layer that try to recognize the relation between each pair of
source words by concatenating their representation vectors. By concatenating the convolutional representation of some word \( w \) with those of all other ones, the word \( w \) is related to a set of vectors; Then, through another Multi-Layer Perceptron (MLP) layer, the word \( w \) will be represented by another set of vectors, which might contain the information of relations with all other words in the sentence except this source word; At last, we sum these vectors together to generate the output for word \( w \) (see Figure 3).

**MLP\(_{DEC} \)**: The last module of the relation layer called MLP decoder changes the representation dimension for each source word and get the final representation of each source word which contains the relations information from all other words, we call it as **entangled sequential representation**. Accordingly, for convenient distinction, we name the initial representation generated by NMT encoder as **recurrent sequential representation**.

As shown in Figure 3, the entangled sequential representations together with recurrent sequential representations of all source words are fed into the attention layer of NMT decoder, we update the Formula (5) here

\[
g(s_{j-1}, h_i) = v^T_a \tanh(W_a s_{t-1} + U_a h_i + r_i) \tag{9}
\]

where \( r_i \) is the entangled sequential representation of each source word \( x_i \) created by relation layer. \( r_i \) is calculated by taking the i-th vector of matrix \( r \)

\[
r = MLP_{DEC}(GPL(CENC(h))) \tag{10}
\]

As noted in Formula (10), the recurrent sequential representations are encoded by the convolutional layer to be fixed-dimension vectors, then GPL is adopted to collect information propagated between each vectors pair and feed new vectors \( \overline{h}_i \) to the MLP decoder layer to recover vector dimension back. \( \overline{h}_i \) is the sum of the i-th column vectors produced by GPL layer as shown in Formula (11), MLP\(_{DEC} \) is actually a multi-layer fully connected network.

\[
\overline{h}_i = \sum_j GPL(h_i, h_j) \tag{11}
\]

Residual connections is adopted from recurrent sequential representation to entangled sequential representation by addition, multiplication or concatenation operation to generate final output.

**Convolution Gated Recurrent Unit**

Here, we introduce a Convolution Gated Recurrent Unit (CGRU) (Kaiser and Sutskever 2015) to replace the 1-dimensional convolutional encoder layer in Figure 3, CGRU is a unit which is similar with the standard GRU but combines the convolution operation into GRU (Cho et al. 2014b), with some following changes, the state \( s_{t-1} \) is updated into \( s_t \) layer by layer:

\[
s_t = u_t \odot s_{t-1} + (1 - u_t) \odot c_t
\]

\[
c_t = \tanh(W_c \ast (r_t \odot s_{t-1}) + b_c)
\]

\[
u_t = \sigma(W_u \ast s_{t-1} + b_u)
\]

\[
r_t = \sigma(W_r \ast s_{t-1} + b_r)
\]

where \( W \) and \( b \) can be treated as the weights and bias of convolutional operation, they are learnable parameters. Such as the operation \( W_u \ast s_{t-1} \) represents a convolution operation on the state \( s_{t-1} \), \( \odot \) denotes the element-wise multiplication, and \( \sigma \) is the sigmoid activation function.

We treat the recurrent operation among layers, without recurrent unit inside one layer, recurrent sequential representation is feed into one CGRU layer to produce the next state, which is actually the input of next CGRU layer; we apply the CGRU to the recurrent sequential representation for several
times, and the final output is used as the representation of the source sentence, feed into the graph propagation layer.

Experiments

Data Preparation
We conduct experiments on Chinese-English translation task by using two different scale data sets, one of which is a smaller data set, and the other is larger one.

IWSLT Data The training data set of the IWSLT corpus consists of 44K sentences from the tourism and travel domain. The development data set was composed of the ASR devset 1 and devset 2 from IWSLT 2005, and the test data set is the IWSLT 2005 test set.

NIST Data The training data is consisted of 1M Chinese-English parallel sentence pairs with 19M source tokens and 24M target tokens from the LDC corpora of LDC2002E18, LDC2003E07, LDC2003E14, and Hansard’s portion of LDC2004T07, LDC2004T08 and LDC2005T06. and we use NIST 2002 test data set as validation data set, and other four NIST test data sets are selected as the test sets: NIST 2003 (919 sentences), NIST 2004 (1788 sentences), NIST 2005 (1082 sentences) and NIST 2006 (1664 sentences).

Systems
We compare the translation performance of our system with four other baseline translation systems:

- Traditional statistic machine translation (SMT) system (MOSES)
- The RNNsearch-based translation system (named GROUNDHOG) we reimplement based on (Bahdanau, Cho, and Bengio 2014), this model is conducted as our weaker baseline model.
- The improved RNNsearch-based translation system (denoted by RNNSEARCH*) described above, which we use as a stronger baseline model.
- Google’s new neural translation system (indicated by TRANSFORMER) without RNN units (Vaswani et al. 2017).

Evaluation Mitrces
Without UNK replacement and de-tokenization, the translation is evaluated by using case-insensitive 4-gram BLEU score (Papineni et al. 2002) with test of statistical significance (Collins, Koen, and Kučerová 2005) between the proposed model and baseline models.

Improvement to Translation Performance
To validate the stability and effectiveness of proposed model, we compare it with others on two different scale data sets according to the BLEU score. Firstly, two NMT baseline models do not have advantages comparing with MOSES on small training data, achieving similar BLEU score with MOSES on the test data, which we could get from Tabel 1; On the toy data set, we stably improve BLEU score by using kinds of different relation network configurations, our model with best setting significantly outperforms MOSES by 4.05 points and RNNSEARCH* by 4.02 points. Models with other settings all produce more than 1 points than three baseline models.

On the 1M training data set, RNNSEARCH* produces 3.47+ and 4.12+ BLEU score averagely comparing with MOSES and GROUNDHOG which shows that RNNSEARCH* is a strong baseline model; Besides, it is consistent with Google’s paper that TRANSFORMER produces similar results with RNNSEARCH* on our training data set. Table 2 shows the final results of the systems, our model with best configurations outperforms the strong NMT baseline model RNNSEARCH* by 1.13 points on average, achieving statistically significant improvements on three testing data sets; Moreover, all other configurations in the Table 2 enable our model to improve translation performance in contrast to the strong baseline model.

Linguistic Analysis
To explore the reason why translation performance becomes better, we analyse from three following aspects:

Translation results Table 3 shows several translation samples produced by the NMT strong baseline model and the proposed model. By comparison, from the boldfaced section of our translation result, we could conclude that the NMT baseline model often miss some information of the source sentence (we call under-translation), such as the baseline translation of the first example do not produce the information about when i arrive at the intersection, RNNSEARCH* model loses the information about haircut when generating the target text for the second sample; It similarly happens that, for the sixth one, baseline model fails to capture the latter clause with adversative relation and so on; Besides, another phenomenon we can get is that the longer the source sentence is, it is easier to ignore important informations for the baseline model; While our model could capture all the information source sentence contains successfully.

Specifically, Whether RNN in the encoder and decoder of baseline model or the CNN are both weak to capture the long dependency information, RNNs are skilled in modeling the order information of the sequence, while CNNs mainly focus on local features around some specific word. However, facts prove that proposed relation layer integrates several CNN layers into Bidirectional RNN or based on Bi-RNN, which alleviates the both disadvantages effectively.

Word Alignment Along with the translation results, we also produce the word alignment matrix based on each target word’s attention probability distribution over the whole source sentence for each decoding step. We randomly sample two source sentences from the testing data set, and output both of their alignment matrices on the baseline model RNNSEARCH* and the proposed model RNNSEARCH*+RN.

As shown in Figure 5, for the first example, from the view of source side, the source chinese word yi is contributed to generate three english words the, is and for, which obviously do not make sense, knowledges in the grammar level show that the word yi is only aligned to the english word for, just like the result of our model; Beside, on the target’s ground,
one single word in the English phrase should be aligned to specific source word together with other words in the phrase, such as English "new" and "is:" For the second sentence, unlike the baseline model, our model produces the correct translation "jazz music for jueshi yinyue" and the right alignment, the together with "origin" is aligned to the source word, while baseline model mistakenly aligns the to two source words almost equally.

The baseline RNNsearch* model only use RNN to model the representation of the source sentence and the generation of the target words. Thus, it may ignore the long-distance information; More importantly, because bidirectional RNN models the source sentence sequentially, so, each encoded source word contains the information of all words before it and after it, the generation of one specific target word may be misled by some irrelevant or translated source words and target words (Tu et al.). While our relation network make more explicit guidance for the generation of each decoding step, thereby producing more unambiguous word alignment.

Relationship of Translation with Source Length the performance of the conventional RNN Encoder–Decoder (Cho et al. 2014b) dramatically drops as the length of the sentences increases, and the RNNsearch model are already more robust to the length of the sentences; However, our proposed model behaves stronger for the long source sentences than RNNsearch model (shown in Figure 4).

In the figure, we compare our model with RNNsearch* and Transformer to observe which one have an advan-

| Systems                          | Tuning | Test  | †  |
|----------------------------------|--------|-------|----|
| MOSES                            | 48.49  | 52.35 | -  |
| GROUNDHOG                        | 48.49  | 52.35 | -  |
| RNNsearch*                       | 48.49  | 52.35 | -  |
| +RN x 1 [CENC(Conv+ReLU)+Res(Add)+MLPDEC] | 48.49  | 52.35 | -  |
| +RN x 1 [CENC(Conv+Sigmoid)+Res(Add)+MLPDEC] | 48.49  | 52.35 | -  |
| +RN x 1 [CENC(Conv+LR)+Res(Add)+MLPDEC] | 48.49  | 52.35 | -  |
| +RN x 1 [CENC(Conv+ReLU)+Res(Mul)+MLPDEC] | 48.49  | 52.35 | -  |
| +RN x 1 [CENC(Conv+LR)+Res(Mul)+MLPDEC] | 48.49  | 52.35 | -  |
| +RN x 2 [CENC(Conv+ReLu)+Res(Mul+DC)+MLPDEC] | 48.49  | 52.35 | -  |
| +RN x 2 [CENC(Conv+ReLu)+Res(Mul+DC)+MLPDEC] | 48.49  | 52.35 | -  |
| +RN x 2 [CENC(Conv+LR)+Res(Mul+DC)+MLPDEC] | 48.49  | 52.35 | -  |
| +RN x 2 [CENC(Conv+LR)+Res(Add+DC)+MLPDEC] | 48.49  | 52.35 | -  |
| +RN x 2 [CENC(Conv+[3,5,7]+LR)+Res(Add+DC)+MLPDEC] | 48.49  | 52.35 | -  |

Table 1: Comparison among systems on IWSLT data set, “ReLU”, “Sigmoid” and “LR” stands for the rectified linear, sigmoid and the leaky rectified linear activation function; “Res” represents the residual connection whose operations we used include addition (“Add”) and multiplication (“Mul”); while “DC” is the abbreviation of dense concatenation (Huang et al. 2016). The tuning column indicates the highest BLEU score on the validation set, the bold font shows that the results are better than those of all baseline models, the symbol † and * indicate that proposed model has statistically significant difference (p < 0.05 and p < 0.01 separately) from the baseline system RNNsearch*.

| Systems                          | Tuning | MT 03 | MT 04 | MT 05 | MT 06 | Mean  |
|----------------------------------|--------|-------|-------|-------|-------|-------|
| MOSES                            | 32.31  | 29.96 | 31.21 | 28.61 | 29.10 | 29.72 |
| GROUNDHOG                        | 31.22  | 27.87 | 31.38 | 28.41 | 28.62 | 29.07 |
| TRANSFORMER                      | 34.39  | 32.89 | 35.56 | 31.97 | 32.10 | 33.13 |
| RNNsearch*                       | 34.85  | 32.33 | 35.99 | 32.26 | 32.17 | 33.19 |
| +RN x 2 [CENC(Conv+LR)+Res(Add+DC)+MLPDEC] | 35.53  | 33.38 | 36.64 | 32.67 | 33.10 | 33.95 ± 0.76 |
| +RN x 1 [CENC(Conv+LR)+Res(Add+DC)+MLPDEC] | 35.22  | 33.26 | 36.80 | 33.04 | 33.12 | 34.06 ± 0.87 |
| +RN x 2 [CENC(Conv+[3,5,7]+LR)+Res(Add+DC)+MLPDEC] | 35.01  | 33.29 | 36.42 | 33.11 | 33.09 | 33.98 ± 0.79 |
| +RN x 1 [CENC(Conv+[3,5,7]+LR)+Res(Add+DC)+MLPDEC] | 36.17  | 33.67 | 36.5  | 33.50 | 33.54 | 34.30 ± 1.11 |
| +RN x 2 [CENC(Conv+[3,5,7]+LR)+Res(Add+DC)+MLPDEC] | 35.49  | 33.75 | 36.87 | 33.26 | 33.38 | 34.32 ± 1.13 |

Table 2: Comparison among systems on 1M training data set, meanings of all symbols are the same with Table 1 above except that BLEU scores in the “Mean” column are the arithmetic means of those on the corresponding four test sets.

Our proposed model add an extra relation network layer on the top of the standard bi-directional RNN encoder in the NMT model; Experiments show that our model stably improves the translation performance on both toy and large-scale data sets; Analysis indicates that relation network layer makes the attention mechanism more specific, meanwhile without losing the universality of modeling the order of sequence, which means our model can learn which section should be attended in the process of generating each target word more precisely, simultaneously, capable of keeping the important source information for the long source sentence.

**Conclusion**

We introduce a relation network layer on the top of the standard bi-directional RNN encoder in the NMT model; Experiments show that our model stably improves the translation performance on both toy and large-scale data sets; Analysis indicates that relation network layer makes the attention mechanism more specific, meanwhile without losing the universality of modeling the order of sequence, which means our model can learn which section should be attended in the process of generating each target word more precisely, simultaneously, capable of keeping the important source information for the long source sentence.

**References**

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural Machine Translation By Jointly Learning To Align and Translate. Iclr 2015 1–15.

Bastings, J.; Titov, I.; Aziz, W.; Marcheggiani, D.; and Siman ‘an, K. 2017. Graph Convolutional Encoders for Syntax-aware Neural Machine Translation. Emnlp.

Cho, K.; van Merrienboer, B.; Bahdanau, D.; and Bengio, Y. 2014a. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation 103–111.

Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014b. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) 1724–1734.

Collins, M.; Koehn, P.; and Živković, I. 2005. Clause restructuring for statistical machine translation. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05 (June):531–540.

Heafield, K.; Pouzyrevsky, I.; Clark, J. H.; and Koehn, P. 2013. Scalable Modified Kneser-Ney Language Model Estimation. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 690–696.

Huang, G.; Liu, Z.; Weinberger, K. Q.; and van der Maaten, L. 2016. Densely Connected Convolutional Networks.

Kaiser, Ł., and Sutskever, I. 2015. Neural gpus learn algorithms. arXiv preprint arXiv:1511.08228.

Koehn, P.; Hoang, H.; and Birch, A. 2007. Moses: Open source toolkit for statistical machine translation. Proceedings of the 45th . . . (June):177–180.

Meng, F.; Lu, Z.; Li, H.; and Liu, Q. 2016. Interactive Attention for Neural Machine Translation. Coling-2016 2174–2185.

Och, F. J. 2003. Minimum error rate training in statistical machine translation. Annual Meeting of the ACL 1001(1):160.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. BLEU: a method for automatic evaluation of machine trans-
Figure 5: Word-alignments comparison of proposed model (RNNSEARCH⋆+RN) with RNNSEARCH⋆ baseline model on two samples; (a) represents two alignment results produced by baseline model; (b) indicates the results of our model.

Hyper Parameters and Training Details

- **Moses**: We use the open-source translation system *Moses*¹ (Koehn, Hoang, and Birch 2007) as our SMT baseline; Following detailed configurations are applied to train the SMT model: we train the 4-gram language model on the target side of Large NIST training data by using KenLM open-source toolkit² (Heafield et al. 2013); GIZA++³ (Och 2003) is used to produce the word alignment on the given parallel training data; Default maximal phrase length 7 is employed when we extract the phrase pairs based on word alignment results; Relative translation frequencies and lexical translation probabilities are calculated on both directions, meanwhile, distortion distance and word penalty are also applied; Besides, we set the maximum iterations to be 20 when tuning the parameters, without filtering the phrase table.

- **GROUNDHOG**: We reimplement the attention-based RNNSearch model of (Bahdanau, Cho, and Bengio 2014) by PyTorch deep learning framework ⁴. Concretely, for all NMT baseline systems, we employ a little bit different settings with (Bahdanau, Cho, and Bengio 2014): we filter the training sentence-pairs whose source sentence or target sentence contains more than 50 tokens, the dimension of source and target word embedding is 512, dimensions of all hidden units in both encoder and decoder RNN are all 512, the vocabulary sizes of the source and target side are both 30K. We initialize all weight matrices or vectors by using uniform distribution, all parameters are updated once for each mini-batch by Stochastic Gradient Descent (SGD), the batch size option is 80, in addition, the learning rate is adjusted by optimizer AdaDelta (Zeiler 2012) with the coefficient $\rho = 0.9$ used for computing a running average of squared gradients, and the term $\epsilon = 10^{-6}$ added to the denominator to improve numerical stability. Dropout rate 0.5 is applied in the training process, without dropout and with beam size 10 in decoding. At last, best model on validation set is used to generate translations of testing data.

- **RNNSEARCH⋆**: The decoder of the standard RNNsearch model is changed into the improved one, with all other configurations and training details the same with the GROUNDHOG system above. For leaky ReLU activation function, we use the coefficient 0.1 to control the angle of the negative slope.

- **TRANSFORMER**: We evaluate this model on an open source platform *sockeye*⁵ developed by amazon which is a sequence-to-sequence framework with a focus on NMT based on Apache MXNet, configuration is set by default with both the source and target vocabulary sizes the same as those of above NMT systems. We set layers number to be 4 for both encoder and decoder with no dropout, model hidden size and attention head number are 512 and 8 respectively. During training, smoothed cross entropy loss is adopted and weight tying is used to regularize weight parameters in classification layer.

¹http://www.statmt.org/moses/
²https://github.com/kpu/kenlm
³https://github.com/moses-smt/giza-pp
⁴http://pytorch.org
⁵https://github.com/awslabs/sockeye
Reference in recent years, we have been advocating "integrity" and we want to "forge a government-in-

my light was green when i got to the intersection.

it was in the right with the light in my bag.

it was green in the light when i arrive at the intersection.

Source 我 想 预约一下理发。

Reference i’d like to make an appointment for a haircut.

RNNSEARCH* i’d like to make a reservation.

RNNSEARCH*+RN i’d like to make an appointment for a haircut.

Source 黑夜处处有,神州最漫长。

Reference dark night falls everywhere, but it is the longest in the divine land of china.

RNNSEARCH* in the dark, the land of the motherland is the most long time.

RNNSEARCH*+RN there are everywhere in the sky and the divine land is the longest.

Source 经过国际奥委会的不懈努力, 意大利方面在冬奥会开幕前四天作出让步, 承诺冬奥会期间警方不会进入奥运村搜查运动员驻地, 但是, 药检呈阳性 的运动员仍将接受意大利检察机关的调查。

Reference through the untiring efforts of the ioc, the italian side made concession four days prior to the opening of the international olympic committee.

RNNSEARCH* through the unremitting efforts of the ioc, the italian side made a concession four days prior to the opening of the (UNK) and promised that the police would not be able to search for the athlete ‘ s place during the opening period.

RNNSEARCH*+RN with the unremitting efforts of the international olympic committee, the italian side made a concession in four days before the opening of the (UNK) and promised that the police would not be able to search for the athlete ‘ s place during the opening period.

Source 在最近几年, 我们一直倡导 “诚信”, 要 “打造阳光政府”, 要尊重公众的 “知情权”, 要提高行政 “透明度”, 然而, 事实距离理想还有很大差距。

Reference in recent years, we have been advocating "integrity" and we want to "forge a government-in-

sunshine", improve the "transparency" of government administration, and respect the public’s "right to know". however, the reality is still very far from ideal.

RNNSEARCH* in recent years, we have always advocated " honesty " and " build a sunshine government ", and we must respect the public ' s " right to understand " and to enhance the " transparency " of the " transparency " of the public.

RNNSEARCH*+RN in recent years, we have advocated " integrity " and " build up the sun " . we should respect public " right to know " and improve the " transparency " of the public. however, there is still a big gap between reality and ideals.

Source 是一个内阁关防长官安倍晋三，另外一个是最近屡屡在日的日本外相麻生太郎。

Reference one is chief cabinet secretary shinzo abe, and the other is japanese foreign minister taro aso, who has been made a series of appalling remarks recently.

RNNSEARCH* it is a cabinet minister shinzo abe , another one, that is, recently, the japanese foreign minister taro aso , who has been repeatedly (UNK)

RNNSEARCH*+RN it was the chief cabinet , shinzo abe , and another one that is , er , in recent years, the japanese foreign minister taro aso , who has been repeatedly frequently.

Table 3: Sample Translations