

Deep Neural Network Based Machine Translation System Combination

LONG ZHOU, JIAJUN ZHANG, XIAOMIAN KANG, and CHENGQING ZONG, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, P.R.China

Deep neural networks (DNNs) have provably enhanced the state-of-the-art natural language process (NLP) with their capability of feature learning and representation. As one of the more challenging NLP tasks, neural machine translation (NMT) becomes a new approach to machine translation and generates much more fluent results compared to statistical machine translation (SMT). However, SMT is usually better than NMT in translation adequacy and word coverage. It is therefore a promising direction to combine the advantages of both NMT and SMT. In this paper, we propose a deep neural network based system combination framework leveraging both minimum bayes-risk decoding and multi-source NMT, which take as input the N-best outputs of NMT and SMT systems and produce the final translation. In particular, we apply the proposed model to both RNN and self-attention networks with different segmentation granularity. We verify our approach empirically through a series of experiments on resource-rich Chinese \Rightarrow English and low-resource English \Rightarrow Vietnamese translation tasks. Experimental results demonstrate the effectiveness and universality of our proposed approach, which significantly outperforms the conventional system combination methods and the best individual system output.

CCS Concepts: • **Computing methodologies** \rightarrow **Machine translation**.

Additional Key Words and Phrases: DNN, SMT, NMT, system combination, minimal bayes-risk decoding, low-resource translation

ACM Reference Format:

Long Zhou, Jiajun Zhang, Xiaomian Kang, and Chengqing Zong. 2020. Deep Neural Network Based Machine Translation System Combination. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 1, 1, Article 1 (January 2020), 19 pages. <https://doi.org/10.1145/3389791>

1 INTRODUCTION

Due to the powerful capacity of modeling complex functions and capturing complex linguistic structures, deep neural networks (DNNs) have made big breakthroughs in natural language process (NLP). Specifically, in machine translation (MT), deep learning-based methods have made significant progress in recent years and quickly become the new de facto paradigm of MT in both academia and industry.

In the past seventy years, several paradigms have been developed to solve the MT problem, including phrase-based [22], hierarchical phrase-based [7], RNN-based [41], CNN-based [13], self-attention (Transformer) based [44] approaches. Unlike conventional statistical machine translation

Authors' address: Long Zhou, long.zhou@nlpr.ia.ac.cn; Jiajun Zhang, jjzhang@nlpr.ia.ac.cn; Xiaomian Kang, xiaomian.kang@nlpr.ia.ac.cn; Chengqing Zong, cqzong@nlpr.ia.ac.cn, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, P.R.China, Intelligence Building, No. 95, Zhongguancun East Road, Haidian District, Beijing, 100190.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Association for Computing Machinery.

2375-4699/2020/1-ART1

<https://doi.org/10.1145/3389791>

(SMT) which contains multiple separately tuned components, neural machine translation (NMT) is a single, large neural network which heavily relies on an encoder-decoder framework. Since NMT and SMT systems are two kinds of translation models with large differences, each approach has some advantages and limitations. For example, most sentences in NMT are more fluent than translations by SMT [7, 22], but NMT has a problem to address translation adequacy [21, 43] especially for the rare and unknown words even using subword method [40]. The strength of NMT lies in that the semantic and structural information can be learned by considering global context. At the same time, it suffers from over-translation and under-translation to some extent [43, 52]. Compared to NMT, SMT does not need to limit the vocabulary and can guarantee translation coverage of source sentences. Obviously, the outputs obtained from NMT and SMT may be dissimilar, and they have different strengths and weaknesses, so system combination can be a good choice.

System combination is a method for combining the output of multiple machine translation engines in order to take benefit of the strengths of each of the individual engines. Traditionally, system combination has been explored respectively in sentence-level, phrase-level, and word-level [6, 9, 23]. Among them, word-level combination approaches that adopt a confusion network for decoding have been quite successful [2, 11, 37]. However, these approaches are mainly designed for SMT without considering the features of NMT results. NMT opts to produce diverse words and free word order, which are quite different from SMT. And this will make it hard to construct a consistent confusion network. Furthermore, traditional system combination approaches cannot guarantee the fluency of the final translation results.

We focus on system combination via a deep neural network in this article. We propose a neural system combination (NSC) framework for machine translation, which is adapted from the multi-source NMT model [54] with minimum bayes-risk decoding. Specifically, to address the N-best translation results which have been proven to have more potential than the top-1 output [24, 28, 39], we first introduce minimum bayes-risk decoding to select the translation hypothesis with the lowest bayesian expectation risk for every translation systems. Then, different encoders are employed to model the semantics of the source language input and each previously selected translation produced by different NMT and SMT systems. Finally, we propose four combination modules, which integrate the multiple context vector representations produced by the encoders into the decoder, to generate the final output token by token. It is worth noting that the proposed neural combination method can combine the outputs of any kind of MT systems.

To test the generalization capacity of our model, we apply the proposed model into two representative seq2seq frameworks using the shallow recurrent neural network (RNN) and deep self-attention network. Moreover, we explore system combination performance at different segmentation granularity, such as word and subword [40]. We extensively evaluate the proposed approach on resource-rich Chinese \Rightarrow English (NIST and WMT) and low-resource English \Rightarrow Vietnamese (IWSLT) language pairs.

The neural system combination has been presented in our previous paper [51]. In this article, we make the following significant extensions to our previous work.

- We introduce a non-parameter minimum bayes-risk decoding method to select the best translation from N-best results, which can be reviewed as a reranking approach and is very effective for subsequent system combination.
- In addition to the shallow RNN model, we also apply the NSC framework to the deep self-attention model, which is much more powerful and achieves the state-of-the-art performance in MT. And we introduce and compare several combination strategies for Transformer-based NSC model, which is capable of combining the fluency of NMT and adequacy of SMT.

- We further verify the effectiveness of our methods on subword segmentation granularity and a low-resource translation task. Moreover, we explore the effect of model average, model ensemble, and bidirectional inference combination.

Experimental results demonstrate that our RNN-based NSC model achieves significant improvement by 5.3 BLEU points over the best single system output and 3.4 BLEU points over the state-of-the-art traditional system combination methods. In addition, the extensive experiments on resource-rich and low-resource translation task with different segmentation granularity show that our Transformer-based NSC model achieves significant improvement over the state-of-the-art Transformer.

2 BACKGROUND

2.1 Statistical Machine Translation

Given a source language sentence x , SMT searches through all the sentences y in the target language and finds the one y^* which maximizes the posterior probability $p(y|x)$. This posterior probability is usually decomposed into two parts using the bayes rule as follows:

$$\begin{aligned} y^* &= \underset{y}{\operatorname{argmax}} p(y|x) \\ &= \underset{y}{\operatorname{argmax}} \frac{p(y) \cdot p(x|y)}{p(x)} \\ &= \underset{y}{\operatorname{argmax}} p(y) \cdot p(x|y) \end{aligned} \quad (1)$$

In which, $p(y)$ is called target language model and $p(x|y)$ is named translation model. This kind of decomposition must adhere to rigid probability constraints and cannot make use of other useful translation features. To solve this problem, Och and Ney [33] advocated the use of log-linear models for statistical machine translation to incorporate arbitrary knowledge sources:

$$P(y|x) = \frac{\exp(\sum_{k=1}^K \lambda_k \cdot h_k(x, y))}{\sum_{y'} \exp(\sum_{k=1}^K \lambda_k \cdot h_k(x, y'))} \quad (2)$$

where $h_k(x, y)$ is a set of features, and λ_k is the feature weight corresponding to the k -th feature.

As shown in Figure 1 (a), the translation process of phrase-based SMT can be divided into three steps: (1) segmenting the source sentence into a sequence of phrases, (2) transforming each source phrase to a target phrase, and (3) rearranging target phrases in an order of target language. The concatenation of target phrases forms a target sentence. Meanwhile, hierarchical phrase-based machine translation [7], syntax-based machine translation [48] have been proposed to improve MT quality. However, SMT still suffers from some key questions, such as data sparsity and feature engineering problems.

2.2 Neural Machine Translation

End-to-end neural machine translation [3, 41] aims to directly map natural languages using neural networks. The major difference from conventional SMT is that NMT is capable of learning representations from data, without the need to design features to capture translation regularities manually.

Generally, NMT follows the encoder-decoder framework. The encoder encodes the source language sentence into semantic representations from which the decoder generates the target language sentence word by word from left to right. Figure 1 (b) illustrates an example of RNN-based NMT (RNMT). Given a source language $X = (x_1, x_2, \dots, x_m)$ and a target language sentence

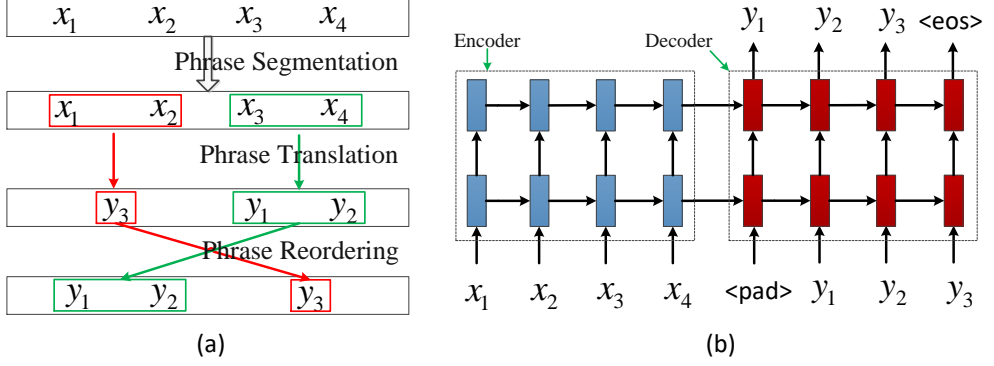


Fig. 1. Translation examples for (a) phrase-based SMT and (b) RNN-based NMT. (x_1, x_2, x_3, x_4) and (y_1, y_2, y_3) are source language and target language, respectively.

$Y = (y_1, y_2, \dots, y_n)$, standard NMT decomposes the sentence-level translation probability as a product of context-dependent word-level translation probabilities:

$$P(y|x; \theta) = \prod_{j=0}^J P(y_j|x, y_{<j}; \theta) \quad (3)$$

The word-level translation probability can be defined as

$$P(y_j|x, y_{<j}; \theta) = \frac{\exp(g(x, y_j, y_{<j}, \theta))}{\sum_y \exp(g(x, y, y_{<j}, \theta))} \quad (4)$$

where $g(x, y_j, y_{<j}, \theta)$ is a real-valued score that indicates how well the j -th target word y_j is given the source context x and target context $y_{<j}$.

There are many design choices in the encoder-decoder framework based on different types of layers, such as RNN-based [41], CNN-based [13], and self-attention based [44] approaches. Particularly, relying entirely on the attention mechanism, the Transformer introduced by [44] can improve the training speed as well as model performance. Although NMT outperforms SMT in terms of BLEU points, NMT still has some drawbacks, such as rare word problems, under-translation, and over-translation.

3 NEURAL SYSTEM COMBINATION FOR MACHINE TRANSLATION

Our goal in this work is to find a way to combine the merits of different translation systems. Macherey and Och [30] gave empirical evidence that these systems to be combined need to be almost uncorrelated in order to be beneficial for system combination. Since NMT and SMT are two kinds of translation models with large differences, we attempt to build a neural system combination model, which can take advantage of the different systems. We will first present our RNN-based system combination model (§3.1) and Transformer-based system combination model (§3.2). Then, we introduce the minimum bayes-risk decoding (§3.3) strategy to select the best translation from top-N hypotheses, which will be sent to our NSC model.

3.1 RNN-based System Combination

Figure 2 illustrates the RNN-based neural system combination framework, which can take as input the source sentence and the results of MT systems. By using bidirectional recurrent neural network

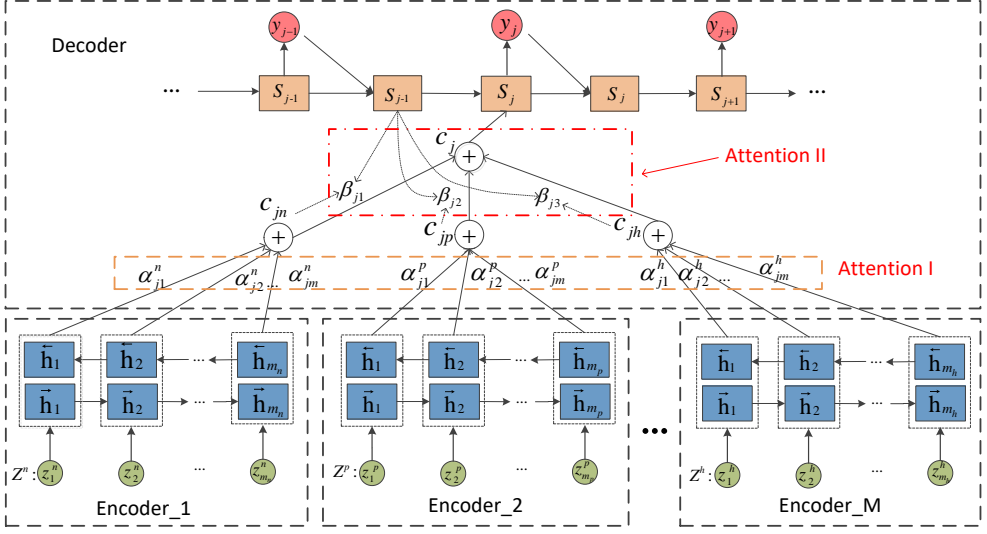


Fig. 2. The architecture of RNN-based system combination model. It is adapted from the multi-source NMT model, which consists of M encoders and one decoder. The hierarchical attention models (Attention I and Attention II) are proposed to access the encoder hidden states of different system outputs.

(Bi-RNN) encoders, translation candidates are encoded to multiple context vector representations, from which the decoder generates the final output word by word. Here, we use MT results as inputs to detail the decoder process.

Formally, the encoders read the translation candidate hypotheses of different MT systems and encode them into a sequence of hidden states $H^k = (h_1^k, h_2^k, \dots, h_m^k)$, and h_i^k is a concatenation of a left-to-right \vec{h}_i^k and a right-to-left \overleftarrow{h}_i^k :

$$h_i^k = \begin{bmatrix} \vec{h}_i^k \\ \overleftarrow{h}_i^k \end{bmatrix} = \begin{bmatrix} f(\vec{h}_{i-1}^k, z_i^k) \\ f(\overleftarrow{h}_{i-1}^k, z_i^k) \end{bmatrix} \quad (5)$$

where $f(\cdot)$ denotes recurrent neural network, such as GRU [8] and LSTM [18], and z^k is the inputs of k -th translation system. Given the hidden states $H(H^1, H^2, \dots, H^K)$ of K MT systems for the same source sentence and previously generated target sequence $Y_{<j} = (y_1, y_2, \dots, y_{j-1})$, the probability of the next target word y_j is

$$p(y_j | Y_{<j}, Z) = \text{softmax}(g(c_j, y_{j-1}, s_j)) \quad (6)$$

Here $g(\cdot)$ is a non-linear function, y_{j-1} represents the word embedding of the previous prediction word, and s_j is the state of decoder at time step j , calculated by

$$s_j = f(\tilde{s}_{j-1}, c_j) \quad (7)$$

$$\tilde{s}_{j-1} = f(s_{j-1}, y_{j-1}) \quad (8)$$

where s_{j-1} is previous hidden state, \tilde{s}_{j-1} is an intermediate state. And c_j is the context vector of system combination obtained by our proposed hierarchical attention mechanism.

Next, we will introduce how to calculate the context vector c_j . Following standard NMT model, we calculate k -th MT system context c_{jk} as a weighted sum of the source annotations:

$$c_{jk} = \sum_{i=1}^m \alpha_{ji}^k h_i \quad (9)$$

where $h_i = [\vec{h}_i; \overleftarrow{h}_i]$ is the annotation of z_i from a bidirectional GRU, and its weight α_{ji}^k is computed by

$$\alpha_{ji}^k = \frac{\exp(e_{ji})}{\sum_{l=1}^m \exp(e_{jl})} \quad (10)$$

where $e_{ji} = v_a^T \tanh(W_a \tilde{s}_{j-1} + U_a h_i)$ scores how well \tilde{s}_{j-1} and h_i match. Equation 9 and Equation 10 are called Attention I in Figure 2.

Given the context vector c_{jk} of different MT outputs, c_j is computed as the weighted sum of the context vectors of different MT systems, just as illustrated in the middle part (Attention II) of Figure 2.

$$c_j = \sum_{k=1}^K \beta_{jk} c_{jk} \quad (11)$$

where K is the number of MT systems, and β_{jk} is a normalized item calculated as follows:

$$\beta_{jk} = \frac{\exp(\tilde{s}_{j-1} \cdot c_{jk})}{\sum_{k'} \exp(\tilde{s}_{j-1} \cdot c_{jk'})} \quad (12)$$

3.2 Transformer-based System Combination

Previous works verified their methods based on shallow recurrent neural network models. However, to obtain state-of-the-art performance, it is essential to leverage recently derived deep models [13, 44], which are much more powerful. Hence, we apply our NSC framework to the deep Transformer model.

For Transformer-based neural system combination model, the neural encoder is identical to that of the dominant Transformer model, which is modeled using the self-attention network [44]. But we use a multi-source framework in which different Transformer encoders are used to capture the semantic of different translation hypotheses. Each encoder is composed of a stack of N identical layers, each of which has two sub-layers:

$$\begin{aligned} \tilde{z}_k^l &= \text{LayerNorm}(z_k^{l-1} + \text{MHAtt}(z_k^{l-1}, z_k^{l-1}, z_k^{l-1})) \\ z_k^l &= \text{LayerNorm}(\tilde{z}_k^l + \text{FFN}(\tilde{z}_k^l)) \end{aligned} \quad (13)$$

where the superscript l indicates layer depth, z_k^l denotes the hidden state of l -th layer of k -th MT system output, FFN means feed-forward networks, and MHAtt denotes the multi-head attention mechanism [44].

For each layer in our decoder, the lowest sub-layer is the masked multi-head self-attention network:

$$s_1^l = \text{LayerNorm}(s^{l-1} + \text{MHAtt}(s^{l-1}, s^{l-1}, s^{l-1})) \quad (14)$$

The second sub-layer is the **combination modules** (ComMod) that integrates sequence contexts of different MT outputs into the decoder:

$$s_2^l = \text{LayerNorm}(s_1^l + \text{ComMod}(s_1^l, z^N)) \quad (15)$$

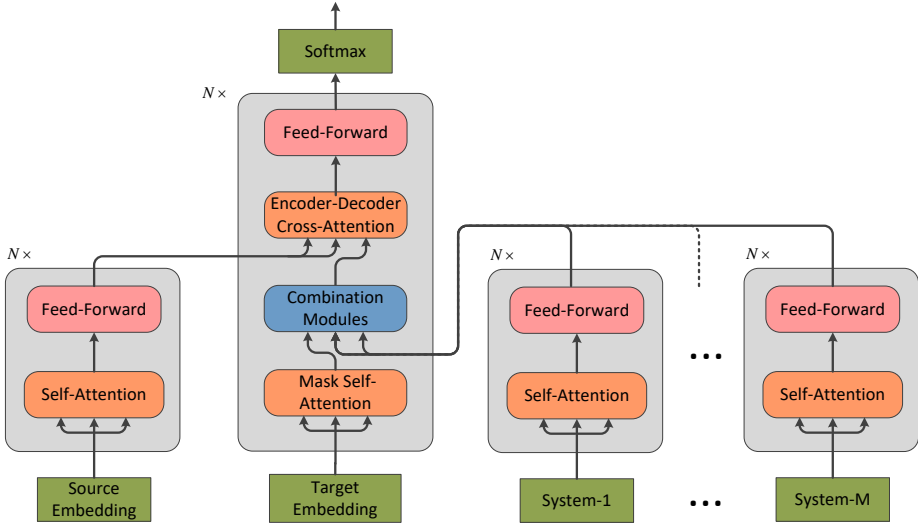


Fig. 3. The extended Transformer translation model that exploits global information produced by additional encoders. Combination modules with blue markers, which uses different strategies to combine the advantages of each system, are the core of the model.

where $z^N = (z_1^N, z_2^N, \dots, z_K^N)$. We will introduce the four combination modules in detail in Section 3.2.1.

Our preliminary experiments show that the source language is critical to the performance of system combination. So, the decoder stacks another two sub-layers to seek task-relevant input semantics to bridge the gap between the input and output language:

$$\begin{aligned} s_3^l &= \text{LayerNorm}(s_2^l + \text{MHAtt}(s_2^l, h^N, h^N)) \\ s^l &= \text{LayerNorm}(s_3^l + \text{FFN}(s_3^l)) \end{aligned} \quad (16)$$

Finally, we use a linear transformation and softmax activation to compute the probability of the next tokens based on s^N :

$$p(y_j | y_{<j}, x, \theta) = \text{softmax}(s^N W) \quad (17)$$

where θ is model parameters and W is the weight matrix.

3.2.1 Combination Modules. The strength of Transformer-based system combination model lies in that the original mask multi-head self-attention and encoder-decoder multi-head cross-attention are preserved, and the newly added model is able to make full use of strengths of multiple MT hypotheses. Inspired by [26, 27], we introduce four strategies for the combination modules: (a) serial; (b) parallel; (c) flat; (d) hierarchical, as illustrated in Figure 4.

(a) The serial module computes the cross-attention one by one for each input encoder. The query set of each subsequent cross-attention is the output of the preceding sub-layer. (b) The parallel module attends to each encoder independently and then sums up the context vectors. (c) The flat module uses all the states of all input encoders as a single set of keys and values. Thus, the attention models a joint distribution over a flattened set of all encoder states. (d) The hierarchical module first computes the attention independently over each input. The resulting contexts are then treated as states of another inputs and the attention is calculated once again over these states.

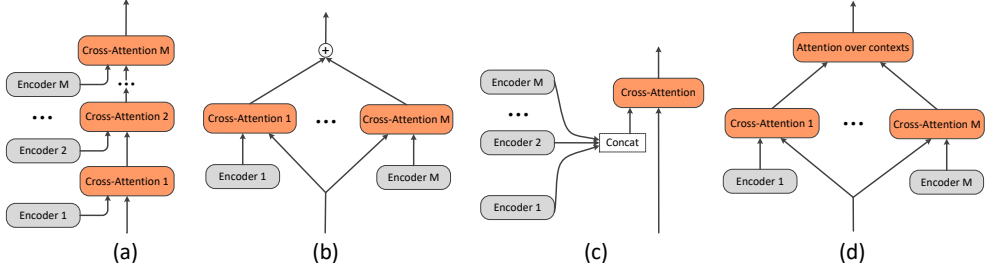


Fig. 4. Four combination modules: (a) serial; (b) parallel; (c) flat; (d) hierarchical. We omit the residual connection and layer normalization in each sub-layer for simplicity.

3.3 Minimum Bayes-Risk Decoding

In the above sections, we propose a neural system combination framework to combine the outputs of different MT systems. However, one MT system may produce N-best translation hypotheses, which have much more potential than 1-best hypothesis but are difficult to employ the above models. Because each translation candidate needs an encoder, this will lead to a more complex and inefficient model. To address this problem, inspired by previous work [23, 42], we propose a minimum bayes-risk decoding strategy, whose role is to select the best one from N-best hypotheses for each MT system.

Given a source language sentence, minimum bayes-risk decoding selects the translation hypothesis with the lowest Bayesian expectation risk from the list of translation hypotheses as the final translation of each translation system.

$$E_{mbr} = \operatorname{argmin}_{Y'} R(Y') = \operatorname{argmin}_{Y'} \sum_Y P(Y|X) L(Y, Y') \quad (18)$$

where $R(Y')$ denotes the bayes risk of candidate translation Y' . $P(Y|X)$ is the conditional probability for the source language sentence x to be translated into the target language sentence y . $L(Y, Y')$ is the loss function, when the minimum bayes risk is calculated by using the BLEU score of the automatic evaluation index of translation quality, it can be expressed as:

$$L_{BLEU}(Y, Y') = 1 - BLEU(Y, Y') \quad (19)$$

where $BLEU(Y, Y')$ is a sentence-level BLEU score. It is worth to note that the minimum bayes-risk decoding method does not require additional parameters and can be directly applied to the inference step.

4 EXPERIMENTS

We evaluate our proposed model on the resource-rich Chinese-English (NIST and WMT) and low-resource Vietnamese-English (IWSLT) translation tasks. The evaluation metric is case-insensitive BLEU [35].

4.1 Data Preparation

For NIST Chinese-English translation, we use the training data from LDC containing 2M bilingual sentence pairs¹. We choose NIST 2003 (MT 03) Chinese-English dataset as the validation set, NIST 2004 (MT04), 2005 (MT05), 2006 (MT06) datasets as our test sets. We limit both Chinese and English vocabulary to 30k in our word-level experiments. We also use BPE [40] to encode Chinese and

¹The corpora includes LDC2000T50, LDC2002T01, LDC2002E18, LDC2003E07, LDC2003E14, LDC2003T17 and LDC2004T07.

English respectively in sub-level experiments. We learn 30K merge operations and limit the source and target vocabularies to the most frequent 30K tokens.

For WMT Chinese-English translation, the models were trained using the parallel corpus without UN dataset from WMT 2017², consisting of about 9 million sentence pairs. We use newsdev2017 and newstest2017 as development and test sets, respectively. We segment each word into subword units using BPE, and vocabulary sizes are 50K for Chinese and English.

For Vietnamese-English translation, we use the provided parallel data (133K sentence pairs) from IWSLT 2015³. We use the TED tst2012 as a valid set and report BLEU scores on TED tst2013. Sentences are encoded using BPE, which has a shared vocabulary of about 20K tokens.

The neural system combination framework should be trained on the outputs of multiple translation systems and the gold target translations. In order to keep consistency in training and testing, we design a strategy to simulate the real scenario. We randomly divide the training corpus into two parts, then reciprocally train the MT system on one half and translate the source sentences of the other half into target translations. The MT translations and the gold target reference can be available.

We list all the translation methods which participate in system combination as follows:

- **PBMT**: It is the start-of-the-art phrase-based SMT system. We use its default setting and train a 4-gram language model on the target portion of the bilingual training data.
- **HPMT**: It is a hierarchical phrase-based SMT system, which uses its default configuration as PBMT in Moses. Without losing the strength of PBMT, HPMT uses hierarchical phrases consisted of both words and subphrases, and has a stronger ability of reordering.
- **RNN-based NMT**: RNMT is an attention-based NMT system with the default setting. It adapts the encoder-decoder framework with recurrent neural network as the core component.
- **Self-Attention based NMT**: Transformer has obtained the state-of-the-art performance on machine translation, which predicts target sentence from left to right relying on self-attention mechanism.

4.2 Training Details

Both RNN-based baseline and NSC model are implemented on the open-source toolkit dl4mt⁴, with most default parameter settings kept the same. The dimension of word embedding is set to 500 and the size of the hidden layer is 1000. The network parameters are updated with Adadelta algorithm. Dropout is also applied to the output layer to avoid over-fitting. We adopt beam search with beam size $b=10$ at test time. As to confusion-network-based system Jane [11], we use its default configuration and train a 4-gram language model on target data and 10M Xinhua portion of Gigaword corpus.

We implement our Transformer-based NSC model based on the open-sourced tensor2tensor⁵ toolkit for training and evaluating. For Chinese-English translation task, we use the hyperparameter settings of base Transformer model as [44], whose encoder and decoder both have 6 layers, and 512 dimension sizes, 8 attention-heads, 2048 feed-forward inner-layer dimensions. At such a small scale of Vietnamese-English translation, we opt for small Transformer models with 5 layers, 256 dimensional size, 2 attention-heads, and 1024 inner-layer dimensions. In minimum bayes-risk decoding, we use top-10 translation results for each system. Additionally, we use a single model

²<http://www.statmt.org/wmt17/translation-task.html>

³<https://wit3.fbk.eu/>

⁴<https://github.com/nyu-dl/dl4mt-tutorial>.

⁵<https://github.com/tensorflow/tensor2tensor>.

| System | MT03 | MT04 | MT05 | MT06 | AVE |
|---------------------|--------------|--------------|--------------|--------------|--------------|
| PBMT | 37.47 | 41.20 | 36.41 | 36.03 | 37.78 |
| HPMT | 38.05 | 41.47 | 36.86 | 36.04 | 38.10 |
| RNMT | 37.91 | 38.95 | 36.02 | 36.65 | 37.38 |
| Jane [11] | 39.83 | 42.75 | 38.63 | 39.10 | 40.08 |
| NSC | 40.64 | 44.81 | 38.80 | 38.26 | 40.63 |
| NSC+Source | 42.16 | 45.51 | 40.28 | 39.03 | 41.75 |
| NSC+Ensemble | 41.67 | 45.95 | 40.37 | 39.02 | 41.75 |
| NSC+Source+Ensemble | 43.55 | 47.09 | 42.02 | 41.10 | 43.44 |

Table 1. Translation results (BLEU score) for different machine translation and system combination methods on NIST Chinese-English translation. Jane is a open source machine translation system combination toolkit that uses confusion network decoding. **Best** and **important** results per category are highlighted. All results of our model are significantly better than the single model ($p < 0.01$).

| System | MT03 | MT04 | MT05 | MT06 | AVE |
|-----------|------------|-------------|------------|------------|--------------|
| RNMT | 1086 | 1145 | 1020 | 708 | 989.8 |
| Our Model | 869 | 1023 | 909 | 609 | 852.5 |

Table 2. The number of unknown words in the results of RNMT and our RNN-based NSC model.

obtained by averaging the last 5 checkpoints for model average, and adopt 4 combination models in the ensemble model for both RNN-based and Transformer-based NSC.

4.3 Results on RNN-based System Combination

4.3.1 Main Results. Table 1 shows the translation results of different systems on word level. The outputs of PBMT, HPMT and RNMT are used as the input to the combination framework. We compare our neural combination system with the best individual engine, and the state-of-the-art traditional combination system Jane, which focuses on system combination via confusion network decoding. The BLEU score of the multi-source neural combination model is 2.53 higher than the best single model HPMT. It is worth noting that the source language input gives a further improvement of +1.12 BLEU points. Hence we will also leverage the source language in subsequent experiments.

As listed in Table 1, Jane outperforms the best single MT system by 1.92 BLEU points. However, our neural combination system with source language gets an improvement of +1.67 BLEU points over Jane. Furthermore, when augmenting our neural combination system with ensemble decoding, it leads to another significant boost of +1.69 BLEU points. In brief, the best result of our proposed model obtains significant improvement by 5.34 BLEU points over the best single system and 3.36 BLEU points over the confusion-network-based system Jane, which demonstrates the superiority of neural network based system combination methods.

4.3.2 Rare and Unknown Words Translation. In order to control the computational complexity, NMT has to employ a small vocabulary, and massive rare words outside the vocabulary are all replaced with a single UNK symbol. Moreover, it is difficult for NMT systems to handle rare words, because low-frequency words in training data cannot capture latent translation mappings in the neural network model. However, we do not need to limit the vocabulary in SMT, which is often able to translate rare words in training data. We count the number of unknown words in translation results for original NMT and our neural combination model. As shown in Table 2, the number of

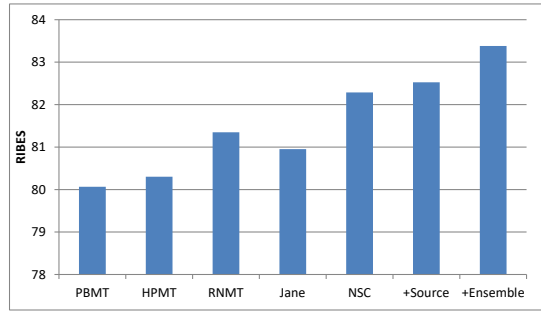


Fig. 5. Translation results (RIBES score) for different machine translation and system combination methods.

| System | MT03 | MT04 | MT05 | MT06 | AVE |
|---------------|--------------|--------------|--------------|--------------|--------------|
| RNMT | 37.91 | 38.95 | 36.02 | 36.65 | 37.38 |
| Enhanced-RNMT | 39.14 | 40.78 | 37.31 | 37.89 | 38.78 |
| Jane [11] | 40.61 | 43.28 | 39.05 | 39.18 | 40.53 |
| Our Model | 43.61 | 47.65 | 42.02 | 41.17 | 43.61 |

Table 3. Translation results (BLEU score) when we replace original NMT with strong Enhance-NMT, which uses ensemble strategy with four NMT models. All results of system combination are based on strong outputs of Enhanced-RNMT.

| # | Strategies | DEV |
|---|--------------------|--------------|
| 1 | PBMT | 37.47 |
| 2 | HPMT | 38.05 |
| 3 | Transformer | 43.41 |
| 4 | NSC (Serial) | 45.51 |
| 5 | NSC (Parallel) | 44.99 |
| 6 | NSC (Flat) | 42.57 |
| 7 | NSC (Hierarchical) | 45.92 |

Table 4. Performance of different combination strategies including serial, parallel, flat, and hierarchical modules.

unknown words of our proposed model is 137 fewer than the original NMT model, which shows the power of our methods in dealing with rare and unknown words.

4.3.3 Translation Fluency. Fluency is the superiority of NMT, and we want to know whether the fluency of combination output has improved. We evaluate fluency (word order) by the automatic evaluation metrics RIBES [19], whose score is a metric based on rank correlation coefficients with word precision. RIBES is known to have a stronger correlation with human judgments than BLEU for English as discussed in [19].

Figure 5 illustrates the experimental results of RIBES scores, which demonstrates that our neural combination model outperforms the best result of the single MT system and Jane. Additionally, although BLEU point of Jane is higher than the single NMT system, the word order of Jane is worse in terms of RIBES. Experiments show that our proposed model can further improve the fluency of NMT.

| Segmentation | System | MT03 | MT04 | MT05 | MT06 | AVE |
|---------------|-----------------------|--------------|--------------|--------------|--------------|--------------|
| word-level | Transformer | 43.41 | 44.34 | 42.63 | 42.88 | 43.31 |
| | +Average | 44.06 | 44.85 | 43.13 | 43.23 | 43.82 |
| | +Ensemble | 44.37 | 45.36 | 43.54 | 43.35 | 44.15 |
| | NSC | 45.92 | 46.95 | 44.06 | 42.43 | 44.84 |
| | +MBR | 46.19 | 48.80 | 45.38 | 43.54 | 45.97 |
| | +Average +Ensemble | 46.52 | 49.15 | 45.88 | 43.89 | 46.36 |
| subword-level | Transformer | 47.82 | 46.57 | 45.29 | 45.66 | 46.33 |
| | +Average | 48.27 | 47.27 | 46.17 | 45.59 | 46.82 |
| | +Ensemble | 48.96 | 47.54 | 46.26 | 46.76 | 47.38 |
| | NSC | 48.87 | 50.44 | 47.87 | 45.55 | 48.18 |
| | +MBR | 49.56 | 50.87 | 47.80 | 46.33 | 48.64 |
| | +Average +Ensemble | 50.19 | 51.34 | 48.44 | 46.76 | 49.18 |
| | | 50.37 | 51.96 | 49.11 | 47.42 | 49.71 |

Table 5. Results for different Transformer-based systems with word-level and subword-level, respectively.

4.3.4 Effect of Ensemble Decoding. In theory, the higher the performance of a single MT system, the higher the performance of system combination. We use ensemble strategy with four NMT models to improve the performance of the original NMT system. As shown in Table 3, the Enhanced-NMT with ensemble strategy outperforms the original NMT system by 1.40 BLEU points, and it has become the best system in all MT systems, which is 0.68 BLEU points higher than HPMT.

After replacing original NMT with strong Enhanced-RNMT, Jane outperforms original result by +0.45 BLEU points, and our model gets an improvement of +3.08 BLEU points over Jane. Experiments further demonstrate that our proposed model is effective and robust for system combination.

4.4 Results on Transformer-based System Combination

4.4.1 Effect of Combination Modules. We first evaluate the proposed four combination strategies including serial, parallel, flat, and hierarchical modules. Table 4 lists the experimental results in Chinese-English development set. Among the combination strategies, the flat strategy is significantly worse than the other three strategies. One possible reason is that concatenating different translation hypotheses results in excessively long source-side encoding, and it is more difficult to handle by encoder-decoder model. In contrast, the hierarchical combination has shown to be the best-performing strategy, and we will use hierarchical strategies in subsequent experiments.

4.4.2 Main Results. The experimental results on NIST Chinese-English translation are depicted in Table 5. In Transformer-based NSC experiments, the single systems involved in the combination are PBMT, HPMT, and Transformer without model average and model ensemble techniques. Compared with other single machine translation systems, Transformer in word-level achieves the best translation quality and significantly outperforms RNMT (Table 1) by +5.93 BLEU points. Based on state-of-the-art Transformer architecture, our NSC model still achieves substantial improvements over the best single model (44.84 vs. 43.31).

4.4.3 Effect on Subword Granularity. Previous work states that neural machine translation models perform particularly poorly on rare words due in part to the smaller vocabularies used by NMT systems. In the past few years, there have been many approaches to deal with rare words and unregistered words. Byte pair encoding (BPE) [40] is the most popular and effective method, which

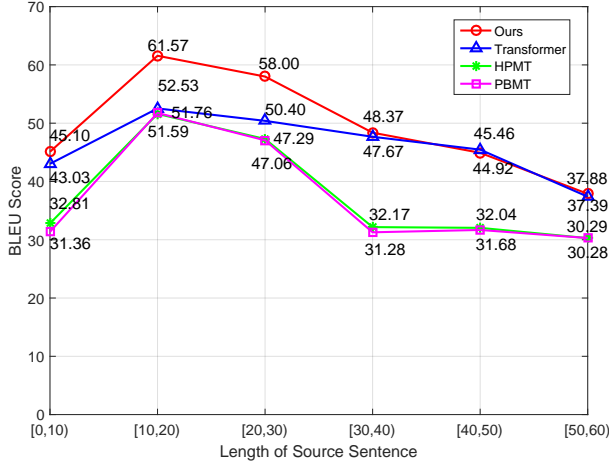


Fig. 6. Length Analysis - performance of translations with respect to the lengths of the source sentences.

is capable of encoding open vocabularies with a compact symbol vocabulary of variable-length subword units.

Hence, we apply BPE to our Transformer baseline and neural system combination models, whose results are shown in Table 5. Although using BPE significantly improves the performance of Transformer baseline (46.33 vs. 43.31), our NSC model in subword-level still obtains an improvement of +1.85 BLEU points than the strong Transformer.

4.4.4 Effect of MBR. The purpose of the MBR is to select the best result from the N-best candidate results of every single system as the inputs of the system combination model. Different from other reranking methods which need more models and parameters to score the candidate hypotheses, our MBR approach only uses N-best results provided by the MT systems, so it requires no additional models or parameters. We perform MBR experiments on both word level and subword level. Experimental results of Table 5 show that MBR leads to another significant boost of +1.13 and +0.46 BLEU points, respectively.

4.4.5 Effect of Model Average and Ensemble. Model average is to average trainable parameters which are saved at last timesteps in a single model, when the model is near convergence. We can get more robust parameters by model average. Besides, model ensemble is a method to integrate the probability distributions of multiple models before predicting the next target word. Model average and model ensemble can also be viewed as a system combination technique that takes advantage of different NSC models.

Table 5 lists the performance of model average and model ensemble in both word level and subword level. Experiments demonstrate the effectiveness of both model average and model ensemble. In particular, two methods of subword level achieve an improvement of +0.54 and +0.53 BLEU points than standard NSC model, respectively.

4.4.6 Length Analysis. Based on the length of source sentences, we divide our test sets into different groups and then compare the system performances in each group. Figure 6 illustrates the BLEU scores on these groups of test sets. NMT (Transformer) performs very well on short source sentences, but degrade on long source sentences. The system combination model has the most obvious improvement in medium-length sentences, where NMT and SMT have similar

| | |
|-------------|--|
| Source | 巴姆大地震受伤的人数大约三万人。 |
| Reference | about 30,000 were injured in the bam earthquake . |
| PBMT | the number of people were injured in the bam earthquake approximately 30,000 . |
| HPMT | the number of bam earthquake wounded about 30,000 . |
| Transformer | about 30,000 people were injured in the baram earthquake . |
| NSC | about 30,000 people were injured in the bam earthquake . |
| Source | 罗斯将这一九九五年事件告诉调查员;该案为美国拳坛舞弊案接受二十个月调查的一部份。 |
| Reference | rose has spoken to investigators about the 1995 incident as part of a 20-month probe of corruption in american boxing circles . |
| PBMT | rose has spoken to investigators about the 1995 incident ; <u>the cases of corruption in american boxing circles as part of a 20 - month probe</u> . |
| HPMT | rose has spoken to investigators about the 1995 incident as part ; <u>the case for the 20 - month probe of corruption in american boxing circles</u> . |
| Transformer | rose told investigators the 1995 incident as part of a 20-month investigation in the us boxing community . |
| NSC | rose has spoken to investigators about the 1995 incident as part of a 20-month probe of <i>corruption</i> in american boxing circles . |

Table 6. Translation examples of single system and our proposed Transformer-based NSC model.

| System | MT03 | MT04 | MT05 | MT06 | AVE |
|-------------------|--------------|--------------|--------------|--------------|--------------|
| Transformer | 47.82 | 46.57 | 45.29 | 45.66 | 46.33 |
| Transformer (R2L) | 45.57 | 45.62 | 44.04 | 44.62 | 44.96 |
| NSC | 48.87 | 50.44 | 47.87 | 45.55 | 48.18 |
| NSC + RNMT - SMT | 48.23 | 47.30 | 45.85 | 46.43 | 46.95 |
| NSC + R2L | 49.88 | 50.97 | 47.74 | 46.24 | 48.71 |

Table 7. Experimental results of ablation study. NSC+RNMT-SMT means that candidate systems of NSC model are Transformer and RNMT. In NSC+R2L, we add a Transformer (R2L) system as a candidate system for system combination.

performance. From the experimental results, we draw the following two conclusions: (1) Because the system combination model still adopts the encoder-decoder framework, it suffers from the quality degradation of long sentences. (2) The system combination model benefits more from candidate systems with similar performance.

4.4.7 Case Study. Table 6 gives two examples to demonstrate the benefits of our proposed NSC model. In the first example, “巴姆” is a rare word for NMT and the baseline NMT (Transformer) incorrectly translates this word into “baram”. With the help of SMT outputs, the NSC model gets the correct translation. In the second example, the source words “舞弊” is untranslated in the output of Transformer. Although PBMT and HPMT translate this word well, they do not conform to the grammar. By combining the merits of NMT and SMT, our model can remedy the errors and obtain high-quality translations in these cases.

4.4.8 Ablation Study. In previous sections, PBMT, HPMT and Transformer are used as combinatorial systems. In this section, we will test the performance of only using RNMT and Transformer

| System | Chinese-English | | Vietnamese-English | |
|-------------|-----------------|--------------|--------------------|--------------|
| | DEV | TEST | DEV | TEST |
| PBMT | 14.44 | 16.33 | 21.73 | 23.66 |
| HPMT | 15.14 | 16.75 | 18.17 | 24.11 |
| Transformer | 22.17 | 23.56 | 21.98 | 24.37 |
| Our Model | 23.37 | 24.49 | 23.11 | 26.08 |

Table 8. Translation performance for large-scale WMT17 Chinese-English and low-resource IWSLT Vietnamese-English language pairs.

outputs, without SMT. Besides, previous work has shown that it is beneficial to integrate bidirectional decoding into NMT [52]. Here we also want to explore whether adding reverse translation as a candidate system can improve the performance of system combination.

The penultimate row of Table 7 shows the experimental results of only using RNMT and Transformer as inputs, from which we can draw the following conclusions: (1) SMT outputs are really useful to NSC model, because NSC with SMT significantly outperforms NSC without SMT; (2) The improvement of BLUE (46.95 vs. 46.33) demonstrates that our NSC framework can combine any outputs and work even without SMT. (3) The greater the diversity of candidate systems, the better the effect of system combination.

To further verify this conclusion, we first train a Transformer model that generates translation in a right-to-left direction. Then we regard it as a candidate system for system combination model. Table 7 shows the translation result. Although Transformer (R2L) performs worse than Transformer, NSC model with R2L obtains +0.53 BLEU points than NSC model without R2L, which shows the effectiveness of bidirectional inference combination.

4.4.9 Results on Large-Scale Language Pairs. We further demonstrate the effectiveness of our model in large-scale WMT17 Chinese-English translation tasks. As listed in Table 8, our approach still significantly outperforms the state-of-the-art Transformer model in development and test sets by 1.20 and 0.93 BLEU points, respectively. Note that NIST Chinese-English datasets contain four reference translations for each source sentence while the WMT Chinese-English datasets only have a single reference.

4.4.10 Results on Low-resource Language Pairs. A well-known property of statistical systems is that increasing amounts of training data lead to better results. We perform additional experiments on a low-resource language pair to better evaluate our system combination model. Table 8 lists the translation performance of Vietnamese-English. Unlike the results in resource-rich Chinese-English translation in which Transformer significantly outperforms PBMT and HPMT, the quality gap between SMT and NMT has become smaller in low-resource Vietnamese-English translation. As shown in Table 8, our proposed NSC model obtains higher BLEU scores compared to the best single system (26.08 vs. 24.37), which demonstrates the universality of neural system combination method.

5 RELATED WORK

Neural machine translation has drawn much attention due to its promising translation performance recently [3, 13, 41, 44, 46]. Most NMT methods are based on the encoder-decoder architecture, which can achieve promising translation performance in a variety of language pairs [20], and rapid adoption in deployments by, e.g., Baidu [50], Google [46], and Microsoft [15].

One branch of related work is improving NMT with SMT techniques. Arthur et al. [1] proposed to incorporate discrete translation lexicons into the NMT model. He et al. [16] presented a log-linear

model to integrate SMT features into NMT. Wang et al. [45] proposed a method that incorporates the translations of SMT into NMT with an auxiliary classifier and a gating function. Zhao et al. [49] regarded SMT phrase table as recommendation memory for NMT. The main idea is to add bonus to words worthy of recommendation, so that NMT can make correct predictions. The major difference between our work and these studies is that basic units and techniques of SMT process are employed in the previous work, while in our work the SMT outputs are utilized to improve NMT.

Our goal in this work is to take advantage of NMT and SMT by system combination, which attempts to find consensus translations among different machine translation systems. In past several years, word-level, phrase-level and sentence-level system combination methods were well studied [4, 5, 11, 17, 24, 25, 29, 31, 36, 38, 53], and reported state-of-the-art performances in benchmarks for SMT. Following our previous work [51], we introduce a neural system combination model which combines the advantages of NMT and SMT efficiently. In this current work, we propose minimum bayes-risk decoding to utilize the N-best translation results. And this article conducts more comprehensive experiments and analyses for system combination model. Specifically, besides RNN-based model, we also explore self-attention network based system combination. Moreover, we conduct the experiments at different segmentation granularity including word and sub-word level, and we also verify the effectiveness on low-resource translation.

There are also some work share a somehow similar idea with our work. Niehues et al. [32] used phrase-based SMT to pre-translate the inputs into target translations. Then the NMT system generated the final hypothesis using the pre-translation. Moreover, multi-source MT has been proved to be very effective to combine multiple source languages [10, 12, 26, 34, 54]. Xia et al. [47] and Geng et al. [14] used two-pass and multi-pass decoding for neural machine translation, respectively. Unlike previous works, we adapt the idea of two-pass inference and employ a multi-source encoder-decoder framework for neural system combination.

6 CONCLUSION

In this work, we propose a simple and effective deep neural network based system combination framework for machine translation. The basic idea consists of two steps: 1) address N-best translation candidates by utilizing minimum bayes-risk decoding; 2) take advantage of NMT and SMT by adapting the multi-source encoder-decoder model. The neural system combination method cannot only address the fluency of NMT and the adequacy of SMT, but also can accommodate the N-best translation candidates and the source sentences as input.

We apply the model to both RNN and self-attention networks with different segmentation granularity. Experiments are conducted on resource-rich Chinese-English and low-resource English-Vietnamese translation tasks. Results show that our approaches can combine any system outputs and obtain significant improvements over the best individual system and the state-of-the-art traditional system combination methods.

7 ACKNOWLEDGMENTS

The research work described in this paper has been supported by the Natural Science Foundation of China under Grant No. U1836221 and the Beijing Municipal Science and Technology Project No. Z181100008918017.

REFERENCES

- [1] Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating Discrete Translation Lexicons into Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 1557–1567. <https://doi.org/10.18653/v1/D16-1162>

- [2] Necip Fazil Ayan, Jing Zheng, and Wen Wang. 2008. Improving alignments for better confusion networks for combining machine translation systems. In *Proceedings of COLING 2008*.
- [3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.
- [4] Srinivas Bangalore, German Bordel, and Giuseppe Richardi. 2001. Computing consensus translation from multiple machine translation systems. In *Proceedings of IEEE ASRU*.
- [5] Debajyoty Banik, Asif Ekbal, Pushpak Bhattacharyya, and Siddhartha Bhattacharyya. 2019. Assembling translations from multi-engine machine translation outputs. *Applied Soft Computing* 78 (2019), 230–239.
- [6] Boxing Chen, Min Zhang, Haizhou Li, and Aiti Aw. 2009. A comparative study of hypothesis alignment and its improvement for machine translation system combination. In *Proceedings of ACL 2009*.
- [7] David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL 2005*.
- [8] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of EMNLP 2014*.
- [9] Yang Feng, Yang Liu, Haitao Mi, Qun Liu, and Yajuan Lu. 2009. Lattice-based system combination for statistical machine translation. In *Proceedings of ACL 2009*.
- [10] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of NAACL-HLT 2016*.
- [11] Markus Freitag, Matthias Huck, and Hermann Ney. 2014. Jane: open source machine translation system combination. In *Proceedings of EACL 2014*.
- [12] Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016*.
- [13] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. PMLR, International Convention Centre, Sydney, Australia, 1243–1252. <http://proceedings.mlr.press/v70/gehring17a.html>
- [14] Xinwei Geng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2018. Adaptive multi-pass decoder for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 523–532.
- [15] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv preprint arXiv:1803.05567* (2018).
- [16] Wei He, Zhongjun He, Hua Wu, and Haifeng Wang. 2016. Improved neural machine translation with SMT features. In *Proceedings of AAAI 2016*.
- [17] Kenneth Heafield and Alon Lavie. 2010. Combining machine translation output with open source. In *The Prague Bulletin of Mathematical Linguistics*.
- [18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [19] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of EMNLP 2010*.
- [20] Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. In *Proceedings of IWSLT 2016*.
- [21] Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, Vancouver, 28–39. <https://doi.org/10.18653/v1/W17-3204>
- [22] Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of ACL NAACL 2013*.
- [23] Shankar Kumar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of HLT-NAACL 2004*.
- [24] Maoxi Li, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2009. The CASIA statistical machine translation system for IWSLT 2009. In *Proceedings of IWSLT2009*.
- [25] Maoxi Li and Chengqing Zong. 2008. Word reordering alignment for combination of statistical machine translation systems. In *Proceedings of the International Symposium on Chinese Spoken Language Processing*.
- [26] Jindřich Libovický and Jindřich Helcl. 2017. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 196–202. <https://doi.org/10.18653/v1/P17-2031>
- [27] Jindřich Libovický, Jindřich Helcl, and David Mareček. 2018. Input Combination Strategies for Multi-Source Transformer Decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, Belgium, Brussels, 253–260. <https://www.aclweb.org/anthology/W18-6326>

- [28] Yuchen Liu, Long Zhou, Yining Wang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2018. A Comparable Study on Model Averaging, Ensembling and Reranking in NMT. In *Natural Language Processing and Chinese Computing*, Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan (Eds.). Springer International Publishing, Cham, 299–308. <http://tcci.ccf.org.cn/conference/2018/papers/166.pdf>
- [29] Wei-Yun Ma and Kathleen Mckeown. 2015. System combination for machine translation through paraphrasing. In *Proceedings of EMNLP 2015*.
- [30] Wolfgang Macherey and Franz Josef Och. 2007. An empirical study on computing consensus translations from multiple machine translation systems. In *Proceedings of EMNLP 2007*.
- [31] Benjamin Marie and Atsushi Fujita. 2018. A Smorgasbord of Features to Combine Phrase-Based and Neural Machine Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*. Association for Machine Translation in the Americas, Boston, MA, 111–124. <https://www.aclweb.org/anthology/W18-1811>
- [32] Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. In *Proceedings of COLING 2016*.
- [33] Franz Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL 2002*.
- [34] Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *Proceedings of MT Summit*.
- [35] Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. Bleu: a methof for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.
- [36] Matiss RIKTERS. 2019. Hybrid Machine Translation by Combining Output from Multiple Machine Translation Systems. *Baltic Journal of Modern Computing* 7, 3 (2019), 301–341.
- [37] Antti-Veikko I. Rosti, Spyros Matsoukas, and Richard Schwartz. 2007. Improved word-level system combination for machine translation. In *Proceedings of ACL 2007*.
- [38] Antti-Veikko I. Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2008. Incremental hypothesis alignment for building confusion networks with appplication to machine translation systems combination. In *Proceedings of the Third ACL Workshop on Statistical Machine Translation*.
- [39] Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh’s Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics, 389–399. <https://doi.org/10.18653/v1/W17-4739>
- [40] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL 2016*.
- [41] Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS 2014*.
- [42] Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice Minimum Bayes-Risk Decoding for Statistical Machine Translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, 620–629. <https://www.aclweb.org/anthology/D08-1065>
- [43] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of ACL 2016*.
- [44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 5998–6008. <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>
- [45] Xing Wang, Zhengdong Lu, Zhaopeng Tu, Hang Li, Deyi Xiong, and Min Zhang. 2017. Neural machine translation advised by statistical machine translation. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- [46] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, and et al Mohammad Norouzi. 2016. Google’s neural machine translation system: bridging the gap between human and machine translation. In *arXiv preprint arXiv:1609.08144*.
- [47] Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *Advances in Neural Information Processing Systems*. 1784–1794.
- [48] Kenji Yamada and Kevin Knight. 2001. A Syntax-based Statistical Translation Model. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Toulouse, France, 523–530. <https://doi.org/10.3115/1073012.1073079>
- [49] Yang Zhao, Yining Wang, Jiajun Zhang, and Chengqing Zong. 2018. Phrase table as recommendation memory for neural machine translation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*.
- [50] Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. 2016. Deep recurrent models with fast-forward connections for neural machine translation. *arXiv preprint arXiv:1606.04199* (2016).

- [51] Long Zhou, Wenpeng Hu, Jiajun Zhang, and Chengqing Zong. 2017. Neural System Combination for Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 378–384. <https://doi.org/10.18653/v1/P17-2060>
- [52] Long Zhou, Jiajun Zhang, and Chengqing Zong. 2019. Synchronous Bidirectional Neural Machine Translation. *Transactions of the Association for Computational Linguistics* 7 (March 2019), 91–105. https://doi.org/10.1162/tac1_a_00256
- [53] Junguo Zhu, Muyun Yang, Sheng Li, and Tiejun Zhao. 2016. Sentence-level paraphrasing for machine translation system combination. In *Proceedings of ICYCSEE 2016*.
- [54] Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of NAACL-HLT 2016*.