

Pivot-Based Semantic Splicing for Neural Machine Translation

Di Liu, Conghui Zhu^(✉), Tiejun Zhao, Xiaoxue Wang,
and Muyun Yang

Harbin Institute of Technology, Harbin 150001, China
{Liudi, chzhu, tjzhao, wangxiaoxue, ymy}@mtlab.hit.edu.cn

Abstract. Current neural machine translation (NMT) usually extracts a fixed-length semantic representation for source sentence, and then depends on this representation to generate corresponding target translation. In this paper, we proposed a pivot-based semantic splicing model (PBSSM) to obtain a semantic representation including more translation information for source sentence, thus improving the translation performance of NMT. The spliced semantic representation is derived from source languages of trilingual parallel corpus by the pivot-based NMT. Besides, the proposed PBSSM only depends on one source language to generate its semantic representation during the encoding process. We integrated it into the NMT architecture. Experiments on the English-Japanese translation task show that our model achieves a substantial improvement by up to 22.9% (3.74 BLEU) over the baseline.

Keywords: Neural machine translation · Pivot-based translation · Semantic splicing

1 Introduction

The neural machine translation systems implemented as encoder-decoder network with recurrent neural networks (Mikolov et al. 2010; Rumelhart et al. 1988; Sundermeyer et al. 2012) have achieved impressive performance in many translation tasks (Sutskever et al. 2014; Cho et al. 2014a). Current neural machine translation (NMT) methods usually extract a fixed-length semantic representation for source sentence, and then generate corresponding target translation depending on the representation (Sutskever et al. 2014). Obviously, the semantic representation obtained by the encoder is essential to NMT.

In order to obtain more effective semantic vector, many researchers use multilingual parallel corpus to train a system that consists multiple encoders and multiple decoders (Luong et al. 2015a; Dong et al. 2015; Ando and Zhang 2004; Cohn et al. 2007). Despite their success, these methods center around learning semantic representation depending on multilingual input to the encoder. However, they neglect the equivalent translation information between multilingual inputs. This work shows that the equivalent translation information is beneficial for NMT.

In this paper, we put forward the pivot-based NMT model, which significantly improves the translation quality of English to Japanese. Based on the pivot-based NMT model, we further enrich the semantic representation, proposing a pivot-based semantic

splicing model (PBSSM) that achieves a substantial improvement of up to 3.74 BLEU points over the baseline.

2 Background

In this section, we mainly introduce the pivot-based translation and NMT model with attention mechanism: RNN-Search (Bahdanau et al. 2015). On the basis of these work, we put forward pivot-based NMT model and its semantic splicing extension model (PBSSM).

2.1 Pivot-Based Machine Translation

When being lack of the bilingual parallel corpus from the source language to the target language, the whole translation performance will be degraded. To solve the problem caused by the lack of parallel corpus, the pivot language is introduced. The pivot language as an intermediary establishes a bridge from the source language to the target language. Pivot-based translation model is shown in Fig. 1.

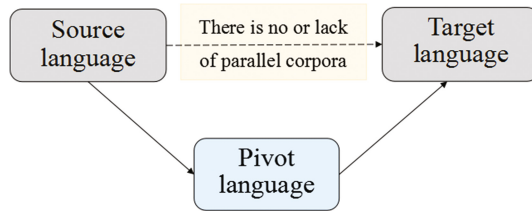


Fig. 1. The Pivot-based translation architecture

The representative research methods of pivot-based translation can be divided into phrase-based translation method (Cohn et al. 2007), sentence-based translation method (Utiyama et al. 2007) and Corpus-based method (Hua et al. 2009). At present, the research on pivot-based translation is mainly carried out in Statistics Machine Translation (SMT). Inspired by the success of NMT, we implement the pivot-based machine translation by neural network, and propose pivot-based NMT model.

2.2 Attention-Based Neural Machine Translation

The basic NMT model consists of an encoder and a decoder. The encoder reads and encodes a source sentence, a sequence of vectors $x = (x_1, x_2, \dots, x_m)$, into a semantic representation C . The decoder then generates one target word y_j , ($1 \leq j \leq n$) at a time from the encoded semantic representation c . Motivated from the observation in (Cho et al. 2014a), Bahdanau adopted attention mechanism in NMT model, proposed the attention-based neural machine translation model (Bahdanau et al. 2015) (Fig. 2).

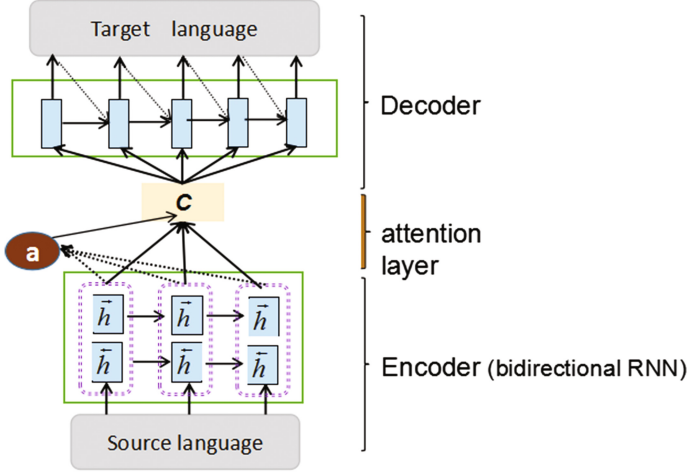


Fig. 2. The attention-based NMT model

The attention mechanism in translation task allows the model to learn to align words when translating. The encoder of this model is constructed by a bidirectional recurrent neural network (BiRNN) (Schuster et al. 1997), which consists of a forward RNN \vec{f} and a reverse RNN \overleftarrow{f} . When the encoder reads an input source sentence $x = (x_1, x_2, \dots, x_m)$, the forward RNN \vec{f} calculates a forward sequence of hidden states $(\vec{h}_1, \dots, \vec{h}_m)$, and the reverse RNN \overleftarrow{f} computes a backward sequence $(\overleftarrow{h}_1, \dots, \overleftarrow{h}_m)$. At each position of the source sentence x , the annotation vector $h_j = [\vec{h}_j, \overleftarrow{h}_j]$ is obtained by concatenating the hidden states \vec{h}_j and \overleftarrow{h}_j . The decoder generates a corresponding translation $y = (y_1, \dots, y_n)$ with Beam-Search algorithm. When given the encoded semantic representation c and all the previously predicted words $y = (y_1, \dots, y_{t-1})$, the decoder uses Eq. (1) to predict the next target word y_i .

$$p(y_i | \{y_1, \dots, y_{i-1}\}, c) = g(y_{i-1}, s_i, c_i) \quad (1)$$

Where g is a nonlinear function that outputs the probability of y_i , and s_i is the hidden state at time i which computed by Eq. (2).

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \quad (2)$$

Where f is a nonlinear function, c_i is related to the hidden states of the input sentence, and is calculated by Eq. (3).

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \quad (3)$$

Where α_{ij} is computed by the following Eq. (4).

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \quad (4)$$

Where $e_{ij} = a(s_i - 1, h_j)$ is an alignment model. It is used to calculate the relevance score, which measures how relevant the j -th encoded semantic representation of the inputs and the output at position i . The score is computed with the decoded hidden state s_i and the j -th hidden state h_j of the encoder. The alignment model a is jointly trained with all other parameters.

3 The Framework of Semantic Splicing Extension

Our baseline is implemented with attention-based neural machine translation model. In order to improve the translation performance of English to Japanese, we utilize multiple parallel corpora to strengthen the representation of source sentence with the method of pivot-based translation. As illustrated in Fig. 3.

In Fig. 3, ① refers to a typical NMT structure, ② is an encoder process, ③ is a decoder process. Referred to in the red dotted line is a typical structure of pivot-based

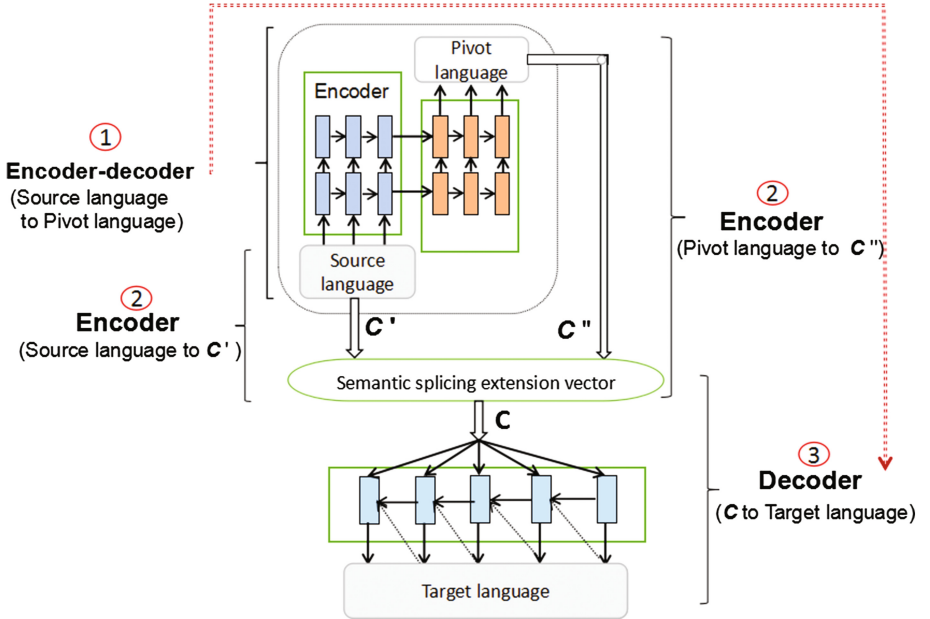


Fig. 3. The framework of semantic splicing extension (Color figure online)

translation, which consists of ① and ②, and ③. It is just the pivot-based NMT model we propose. In addition, to enrich the semantic representation, we put the semantic representation of the source language (② on the left side) and the pivot language (②

on the right side) together to get an extended semantic representation C , and then use c to generate target translation with the decoder (illustrated as ③). This is the other model PBSSM (PBSSM) that we propose.

3.1 Pivot-Based Neural Machine Translation Model

In the case of scarcity of bilingual parallel corpora of source language and target language, the model has a poor performance (Shown in dotted lines in Fig. 1). Based on the research of pivot-based machine translation and neural machine translation, we combine their advantages to improve the translation performance.

Considering the environment of pivot language, we introduce the pivot language between the source language and the target language. Because there exist rich parallel corpora of the source language to the pivot language and the pivot language to the target language, which is crucial to the whole process of translation.

In Fig. 4, *Model_1* and *Model_2* adopt attention NMT model which is described in Sect. 2.2. After we finish training the two separate models, *Model_1* can be used to translate the source language to the pivot language, and then, use *Model_2* to translate the pivot language to the target language. *Model_1* and *Model_2* are two separate models, thus we try to use the available corpus of the source language to the pivot language as much as possible, to improve the translation performance of the source language to the pivot language, so does the whole translation system.

3.2 Pivot-Based Semantic Splicing Model

Most neural machine translation models are trained on bilingual parallel corpora. When we have multilingual parallel corpora, we can make full use of them to improve the performance of translation. Orhan (Orhan et al. 2016) proposes an attention-based encoder-decoder network that admits a shared attention mechanism with multiple encoders and decoders. Orhan proves that using the shared translation system with

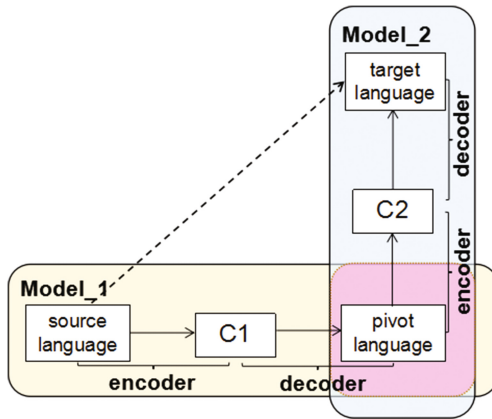


Fig. 4. Pivot-based NMT architecture

multiple parallel corpora can improve the system's performance with less dataset by experiments.

Some of our datasets is trilingual parallel corpora, we can use the semantic similarity between parallel corpora, and treat bilingual parallel corpus as input, which will increase the input information and extend the semantic representation. What described above is shown in Fig. 5.

To extend semantic representation, the system needs bilingual parallel corpora lan_src_1 and lan_src_2 as input. Using the function ϕ to establish a connection between the encoded vector c' from lan_src_1 and the other encoded vector c'' from lan_src_2 ,

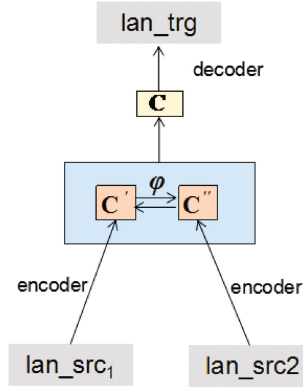


Fig. 5. The structure of extended semantic vector

thus we can get a new vector c from function ϕ that represents the semantic of bilingual source language. Then, use the decoder to generate target language with c . As illustrated in Eq. (3), calculating the semantic representation is associated with hidden states of the encoder. When calculating the hidden states, we create the connection between the hidden state h' of lan_src_1 and the other hidden state h'' of the lan_src_2 , as displayed in Eq. (5) to Eq. (10), where

$$h' = (h'_0, h'_1, \dots, h'_T), h'' = (h''_0, h''_1, \dots, h''_T).$$

where

$$h'_i = [\overrightarrow{h'_i}; \overleftarrow{h'^T_i}]^T, 0 \leq i \leq T_x;$$

$$h''_i = [\overrightarrow{h''_i}; \overleftarrow{h''^T_i}]^T, 0 \leq i \leq T_{x''}.$$

Before the forward hidden states are calculated by the forward RNN, they are randomly initialized with $\overset{I}{h'_0}$; Calculating $\overset{I}{h'_i}, 1 \leq i \leq T_x$ with Eq. (9); Initializing $\overset{I}{h''_0}$

with Eq. (6); Computing $h_i^I, 1 \leq i \leq T_{x''}$ with Eq. (10). And, before the hidden states are calculated by the reverse RNN, they are randomly initialized with $h_{T_{x''}}^S$, Calculating $h_i^S, 1 \leq i \leq T_{x'} - 1$ with Eq. (8); Initializing $h_{T_{x'}}^S$ with Eq. (5); Computing $h_i^S, 1 \leq i \leq T_{x'} - 1$ with Eq. (7).

$$h_{T_{x'}}^S = \sigma(W_{h'x'}^S)x_{T_{x'}}' + W_{h''h'}^S h_{T_0}^S + b_{h'}^S \quad (5)$$

$$h_0^I = \sigma(W_{h''x''}^I)x_0'' + W_{h'h''}^I h_{T_{x''}}^S + b_{h''}^I \quad (6)$$

$$h_i^S = \sigma(W_{h''x''}^S)x_i'' + W_{h''h'}^S h_{i+1}^S + b_{h''}^S \quad (7)$$

$$h_i^S = \sigma(W_{h''x''}^S)x_i'' + W_{h''h'}^S h_{i+1}^S + b_{h''}^S \quad (8)$$

$$h_i^I = \sigma(W_{h''x''}^I)x_i'' + W_{h''h''}^I h_{i-1}^I + b_{h''}^I \quad (9)$$

$$h_i^I = \sigma(W_{h''x''}^I)x_i'' + W_{h''h''}^I h_{i-1}^I + b_{h''}^I \quad (10)$$

where

- x_0' : the first word of lan_src_1 ;
- x_0'' : the first word of lan_src_2 ;
- h_T' : the hidden state of the last word of hidden state h' from lan_src_1 , which is the last component of $h', h' = (h_0', h_1', \dots, h_T')$;
- h_0'' : the component of the zeroth word of the hidden state h'' from lan_src_1 ;
- h_T'' : the hidden state of the last word of hidden state h'' from lan_src_2 , which is the last component of $h'', h'' = (h_0'', h_1'', \dots, h_T'')$;
- σ : non-linear function, generally Sigmoid function or Tanh function;
- w : the corresponding weight matrix;
- b : the corresponding bias vector.

It can be known that the semantic representation c contains the source information of lan_src_1 and lan_src_2 through the analysis of Fig. 5. Therefore, the target language generated by the decoder with c will be more accurate. But, it is not difficult to find that the translation model of lan_src_1 to lan_trg and the other translation model of lan_src_2 to lan_trg are not independent, so do the parameters. And both of the processes of training and testing will need the parallel languages as input. It has a larger price and does not conform to the user's habits. Considering the pivot-based NMT model described in Sect. 3.1, we can extend the semantic representation based on this model,

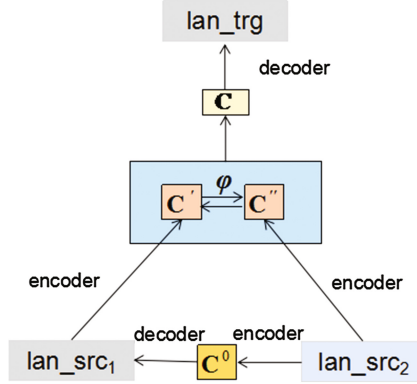


Fig. 6. Pivot-based semantic splicing model

which leads to the PBSSM defined in Fig. 6. Importantly, this model needs bilingual input in training while only monolingual input in testing.

The model with framework shown in Fig. 5 has a disadvantage of requiring bilingual input both training and testing. We can use the advantage of the pivot-based NMT model to make up for this shortcoming. The model we propose is shown in Fig. 6, pivot language is lan_src_1 , source language is lan_src_2 , target language is lan_trg . To enrich semantic representation of pivot language, we still link it with the semantic representation of source language. Therefore, the PBSSM we proposed needs bilingual input in training while only monolingual input in testing.

4 Experiment

In this section, we will first introduce our datasets, parameter settings and experimental results, and then we analyze the results.

4.1 Datasets

We evaluate the proposed model on the task of English-Japanese translations, using more than one parallel corpus. The Trilingual corpus of Chinese, Japanese and English we use in experiment is from Harbin Institute of Technology (HIT) (Yang et al. 2006). HIT corpus contains sports, tourism, transportation, catering, and business and other fields, a totality of 59, 600 pairs of parallel sentences.

In addition, we also use the IWSLT2012 English-Chinese parallel corpus that is oral language corpus, a totality of 72,575 pairs of parallel sentences. Apart from this, there is also Foreign Broadcast Information Service (FBIS) parallel corpus. FBIS is a corpus in the news field, contains about 220, 000 pairs of parallel sentences. Table 1 is the statistics of the scale of the datasets used in the experiment.

Table 1. The statistics of datasets

Corpus	Scale(pair)
HIT	57600
IWSLT2012 (English-Chinese)	72575
FBIS	221348

Table 2. The vocabulary size of different corpus

Corpus	Size
HIT(Chinese)	23000
HIT(English)	18000
HIT(Japanese)	16000
HIT(Chinese) + IWSLT(Chinese)	28000
HIT(English) + IWSLT(English)	21000
HIT(Chinese) + FBIS(Chinese)	30000
HIT(English) + FBIS(English)	30000

4.2 Settings

As a baseline, we use the RNN-search model proposed by (Bahdanau et al. 2015). Since our datasets are basically spoken language, involving much shorter sentences, thus we use sentences of length up to 30 symbols. Both of the encoder and decoder have 1500 cells and 1500 dimensional word embeddings. We train each model using stochastic gradient descent (SGD) with Adam (Kingma and Ba 2015) as an adaptive learning rate algorithm. Each SGD update is computed using a minibatch batch of 80 examples. And, we use the beam-search with beam-width 10 to (approximately) find a translation of maximum log-probability. The vocabulary size of each language is set differently by experiments, shown in Table 2. Our baseline only uses HIT corpus, we trained English-Japanese model and Chinese-Japanese model separately. And, we use HIT corpus, IWSLT2012 corpus and FBIS to train pivot-based NMT model and PBSSM with translation English to Japanese task.

4.3 Experimental Results

We present the translation performance measured by BLEU score in Table 3. We use the RNN-search as our baseline, and test the translation of English to Japanese and the translation of Chinese to Japanese separately with HIT corpus. As illustrated in Table 2, the Chinese to Japanese translation results clearly superior to English to Japanese. Due to, in our datasets, that the parallel corpus of English and Japanese is limited, and the parallel corpus of English and Chinese is abundant, i.e. IWSLT2012 and FBIS. Thus, we take Chinese as the pivot language, English as the source language and Japanese as the target language in pivot-based NMT model and PBSSM. Besides, in the process of translating English to Chinese, we join other corpus, such as IWSLT2012 corpus and FBIS, to improve the translation performance. According to the datasets, we set up three groups of experiments. Among them, we translate Chinese to Japanese only use HIT corpus. While there are three groups settings when translating English to Chinese, they

Table 3. The experimental results

System	Dataset	BLEU(%)
Chinese-Japanese (RNN-Search)	HIT	22.19
English-Japanese (RNN-Search)	HIT	16.35
English-Chinese-Japanese (Pivot-based NMT model)	HIT	14.64
English-Chinese-Japanese (Pivot-based NMT model)	HIT + IWSLT	17.06
English-Chinese-Japanese (Pivot-based NMT model)	HIT + FBIS	19.53
English-Chinese-Japanese (PBSSM)	HIT	15.53
English-Chinese-Japanese (PBSSM)	HIT + IWSLT	17.38
English-Chinese-Japanese (PBSSM)	HIT + FBIS	20.09

are HIT corpus, the union of HIT corpus and IWSLT2012 and the union of HIT corpus and FBIS. We set the parameters as Table 2 and described in Sect. 4.2, use the model introduced in Sect. 3, finish our experiments, the result are shown in Table 3.

4.4 Analysis

In Table 3, the Chinese to Japanese translation result is much better than English to Japanese in our baseline with HIT corpus. This is because the lexical and grammatical structure of Chinese and Japanese is more close than that of English and Japanese. When we only use HIT corpus to train the pivot-based NMT model, the result is worse than the baseline. It is due to the poor performance in English to Chinese translation. When we enrich the parallel corpora of English and Chinese, with the improvement of the translation quality of English to Chinese, the translation result of Chinese to Japanese is also increased. Especially using FBIS to extend the corpus of English to Chinese, English to Japanese translation quality has been greatly improved, due to the large amount of FBIS, which nicely proves the validity of the pivot-based NMT model. And when we use the PBSSM with the same settings and data as the pivot-based NMT model, the translation quality has improved further, which proves the effectiveness of our proposed method once again. It is because the PBSSM uses the extended semantic representation linked with the semantic representation of pivot language which strengthens the information of the encoded semantic representation of the input, thus brings a good experimental result, and the model only requires one language as input in practical use.

5 Conclusion

In this paper, we explore the pivot-based semantic splicing to improve semantic representation of source input in the end-to-end neural machine architecture. We implement the pivot-based translation method by neural network, and combine other related semantic representation to the encoder on multiple parallel corpora. Experiments on the English-Japanese translation task show that our proposed model substantially improves the translation performance.

In the future, we can try to apply more complex splicing function on PBSSM, to get better expression of inputs. In addition, we can add constraint on the encoded semantic expression of input languages.

References

- Mikolov, T., Karafiat, M., Burget, L., Cernock, J., Khudanpur, S.: Recurrent neural network based language model. In: INTERSPEECH, pp. 1045–1048 (2010)
- Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. *Cogn. Model.* **5**(3), 1 (1988)
- Sundermeyer, M., Schlüter, R., Ney, H.: LSTM neural networks for language modeling. *Interspeech*. (2012)
- Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS (2014)
- Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder – Decoder approaches. In: Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, October 2014a
- Luong, M.-T., Le, Q.V., Sutskever, I., Vinyals, O., Kaiser, L.: Multi-task sequence to sequence learning (2015a). arXiv preprint [arXiv:1511.06114](https://arxiv.org/abs/1511.06114)
- Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. *ACL*
- Cho, K., Van Merriënboer, B., Gulcehre, C., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint:1406.1078 (2014)
- Ando, R.K., Zhang, T.: A framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. Technical report RC23462, IBM T.J. Watson Research Center (2004)
- Cohn, T., Lapata, M.: Machine translation by triangulation: Making effective use of multi-parallel corpora. In: *Proceedings ACL* (2007)
- Hua, W., Wang, H.: Revisiting pivot language approach for machine translation. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, vol. 1. Association for Computational Linguistics (2009)
- Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: *ICLR* (2015)
- Utiyama, M., Isahara, H.: A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. *HLT-NAACL* (2007)
- Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**(11), 2673–2681 (1997)
- Boulanger-Lewandowski, N., Bengio, Y., Vincent, P.: Audio Chord Recognition with Recurrent Neural Networks. *ISMIR* (2013)
- Orhan, F., Cho, K., Bengio, Y.: Multi-way, multilingual neural machine translation with a shared attention mechanism (2016). arXiv preprint [arXiv:1601.01073](https://arxiv.org/abs/1601.01073)
- Yang, M., Jiang, H., Zhao, T., Li, S.: Construct Trilingual Parallel Corpus on Demand. In: Huo, Q., Ma, B., Chng, E.-S., Li, H. (eds.) *ISCSLP 2006. LNCS (LNAI)*, vol. 4274, pp. 760–767. Springer, Heidelberg (2006). doi:[10.1007/11939993_76](https://doi.org/10.1007/11939993_76)
- Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: *The International Conference on Learning Representations (ICLR)* (2015)

Machine Translation

12th China Workshop, CWMT 2016, Urumqi, China,

August 25–26, 2016, Revised Selected Papers

Yang, M.; Liu, S. (Eds.)

2016, XI, 125 p. 37 illus., Softcover

ISBN: 978-981-10-3634-7