

Korean-to-Chinese Machine Translation using Chinese Character as Pivot Clue

Jeonghyeok Park^{1,2,3} and Hai Zhao^{1,2,3,*}

¹Department of Computer Science and Engineering, Shanghai Jiao Tong University

²Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, China

³ MoE Key Lab of Artificial Intelligence AI Institute, Shanghai Jiao Tong University
117033990011@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

Korean-Chinese is a low resource language pair, but Korean and Chinese have a lot in common in terms of vocabulary. Sino-Korean words, which can be converted into corresponding Chinese characters, account for more than fifty of the entire Korean vocabulary. Motivated by this, we propose a simple linguistically motivated solution to improve the performance of Korean-to-Chinese neural machine translation model by using their common vocabulary. We adopt Chinese characters as a translation pivot by converting Sino-Korean words in Korean sentence to Chinese characters and then train machine translation model with the converted Korean sentences as source sentences. The experimental results on Korean-to-Chinese translation demonstrate that the models with the proposed method improve translation quality up to 1.5 BLEU points in comparison to the baseline models.

1 Introduction

Neural machine translation (NMT) using sequence-to-sequence structure has achieved remarkable performance for most language pairs (Bahdanau et al., 2014; Cho et al., 2014; Sutskever et al., 2014; Luong and Manning, 2015). Many studies on NMT have tried to improve the translation performance by changing the structure of the network model or adding new strategies (Wu and Zhao, 2018; Zhang

et al., 2018; Xiao et al., 2019). Meanwhile, there are few attempts to improve the performance of the NMT model using linguistic characteristics for several language pairs (Sennrich and Haddow, 2016). On the other hand, Most of the recently proposed statistical machine translation (SMT) systems have attempted to improve translation performance by using linguistic features including part-of-speech (POS) tags (Ueffing and Ney, 2013), syntax (Zhang et al., 2007), semantics (Rafael and Marta, 2011), reordering information (Zang et al., 2015; Zhang et al., 2016) and so on.

In this work, we focus on machine translation between Korean and Chinese, which have few parallel corpora but share a well-known culture heritage, the Sino-Korean words. Chinese loanwords used in Korean are called Sino-Korean words, and can also be written in Chinese characters which are still used by modern Chinese people. Such a shared vocabulary makes the two languages closer despite their huge linguistic difference and provides the possibility for better machine translation.

Because of its long history of contact with China, Koreans have used Chinese characters as their writing system, and even after adopting Hangul(한글 in Korean) as the standard language, Chinese characters have a considerable influence in Korean vocabulary. Currently, the writing system adopted by modern Korean is Hangul, but Chinese characters continue to be used in Korean and Chinese characters used in Korean are called "Hanja". Korean vocabulary can be categorized into native Korean words, Sino-Korean words, and loanwords from other languages. The Sino-Korean vocabulary refers to Ko-

* Corresponding author. This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100) and Key Projects of National Natural Science Foundation of China (U1836222 and 61733011).

Systems	Sentences
Korean	명령은 아래와 같이 <u>반포</u> 되었다.
HH-Convert	命令은 아래와 같이 <u>颁布</u> 되었다.
Chinese	命令颁布如下。
English	The command was promulgated as follows.
Korean	양국은 광범한 영역에서의 <u>공동 이익</u> 을 <u>확인</u> 했다.
HH-Convert	两国은 广范한 领域에서의 共同 利益을 确认했다.
Chinese	两国在广泛的领域确认了共同利益。
English	The two countries have confirmed common interests in a wide range of areas.

Table 1: The HH-Convert is Korean sentence converted by Hangul-Hanja conversion of the Hanjaro. The underline denotes Sino-Korean word and its corresponding Chinese characters in Korean sentence and HH-Convert sentence, respectively.

rean words of Chinese origin and can be converted into corresponding Chinese characters, and considerably account for about 57% of Korean vocabulary. Table 1 shows some sentence pairs of Korean and Chinese with the converted Sino-Korean words. In Table 1, some Chinese words are commonly observed between the converted Korean sentence and the Chinese sentence.

In this paper, we present a novel yet straightforward method for better Korean-to-Chinese MT by exploiting the connection of Sino-Korean vocabulary. We convert all Sino-Korean words in Korean sentences into Chinese characters and take the converted Korean sentences as the updated source data for later MT model training. Our method is applied to two types of NMT models, recurrent neural network (RNN) and the Transformer, and shows significant translation performance improvement.

2 Related Work

There have been studies of linguistic annotation, such as dependency label (Wu et al., 2018; Li et al., 2018a; Li et al., 2018b), semantic role labels (Guan et al., 2019; Li et al., 2019) and so on. Sennrich and Haddow (2016) proved that various linguistic features can be valuable for NMT. In this work, we focus on the linguistic connection between Korean and Chinese to improve Korean-to-Chinese NMT.

There are several studies on Korean-Chinese machine translation. For example, Kim et al. (2002) proposed verb-pattern-based Korean-to-Chinese MT system that uses pattern-based knowledge and consistently manages linguistic peculiari-

ties between language pairs to improve MT performance. Li et al. (2009) improved the translation quality for Chinese-to-Korean SMT by using Chinese syntactic reordering for an adequate generation of Korean verbal phrases.

Since Chinese and Korean belong to entirely different language families in terms of typology and genealogy, many studies also tried to analyze sentence structure and word alignment of the two languages and then proposed the specific methods for their concern (Huang and Choi, 2000; Kim et al., 2002; Li et al., 2008). Lu et al. (2015) proposed a method of translating Korean words into Chinese using the Chinese character knowledge.

There are several attempts to exploit the connection between the source language and the target language in machine translation. Kuang et al. (2018) proposed methods to somewhat shorten the distance between the source and target words in NMT model, and thus strengthen their association, through a technique bridging source and target word embeddings. For other low-resource language pairs, using pivot language to overcome the limitation of the insufficient parallel corpus has been a choice (Habash and Hu, 2009; Zahabi et al., 2013; Ahmadian et al., 2017). Chu et al. (2013) bulid a Chinese character mapping table for Japanese, Traditional Chinese, and Simplified Chinese and verified the effectiveness of shared Chinese characters for Chinese-Japanese MT. Zhao et al. (2013) used the Chinese character, a common form of both languages, as a translation bridge in the Vietnamese-Chinese SMT model, and improved the translation quality by con-

北 선전매체 “北美관계도 “南北관계처럼
대전환”
3.1운동 100주년 맞아 장병 어깨에 원색(
原色) 태극기 부착

Table 2: News headlines with Chinese characters. The underline denotes Chinese characters.

verting Vietnamese syllables into Chinese characters with a pre-specified dictionary. Partially motivated by this work, we turn to Korean in terms of NMT models by fully exploiting the shared Sino-Korean vocabulary between Korean and Chinese.

3 Sino-Korean Words and Chinese Characters

Korea belongs to the Chinese cultural sphere, which means that China has historically influenced regions and countries of East Asia. Before the creation of Hangul (*Korean alphabet*), all documents were written in Chinese characters, and Chinese characters were used continuously even after the creation of Hangul.

Today, the standard writing system in Korea is Hangul, and the use of Chinese characters in Korean sentences is rare, but Chinese characters have left a significant influence on Korean vocabulary. About 290,000 (57%) out of the 510,000 words in the *Standard Korean Language Dictionary* published by the *National Institute of Korean Language* belongs to Sino-Korean words, which were originally written in Chinese characters. Some Sino-Korean words do not currently have corresponding Chinese words and their meanings and usage have changed in the process of introduction, but most of them have corresponding Chinese words. In Korean, Sino-Korean words are mainly used as literary or technical vocabulary and are often used in abstraction concepts and technical terms. The names of people and Korea place are mostly composed of Chinese characters, and newspapers and professional books occasionally use both Hangul and Chinese characters to clarify the meaning. Table 2 shows some news headlines that contain Chinese characters from the Korean news.

Since Korean belongs to alphabetic writing systems and is a language that does not have tones like

Chinese, many homophones were created in their vocabulary in the process of translating the Chinese words into their language. Around 35% of the Sino-Korean words registered in the *Standard Korean Language Dictionary* belong to homophones. Thus converting Sino-Korean words into (usually different) Chinese characters will have a similar impact as semantic disambiguation. For example, the Korean word uisa (의사 in Korean) has many homophones and can have several meanings. To clarify the meaning of the word uisa in Korean context, these words are occasionally written in Chinese characters as follows: 医师 (*doctor*), 意思 (*mind*), 义士 (*martyr*), 议事 (*proceedings*).

In addition, There is a difference between Chinese characters (Hanja) used in Korea and Chinese characters used in China. Chinese can be divided into two categories: Traditional Chinese and Simplified Chinese. Chinese characters used in China and Korea are Simplified Chinese and Traditional Chinese, respectively.

4 The Proposed Approach

The proposed approach for Korean-to-Chinese MT has two phases: Hangul-Hanja conversion and NMT model training. We first convert the Sino-Korean words of the Korean input sentences into Chinese characters, and convert the Traditional Chinese characters of the converted Korean input sentences into Simplified Chinese characters to share the common units between source and target vocabulary. Then we train NMT models with the converted Korean sentences as source data and the original Chinese sentences as target data.

For Hangul-Hanja conversion, we use open toolkit Hanjaro that is provided by the *Institute of Traditional Culture*¹. The Hanjaro can accurately convert Sino-Korean words into Chinese characters and is based on open toolkit UTagger (Shin and Ock (2012) in Korean) developed by the *Korean Language Processing Laboratory of Ulsan University*. More specifically, the Hanjaro first obtains tagging information about morpheme, parts of speech (POS) and homophones of a Korean sentence through the Utagger, and converts Sino-Korean words into corresponding Chinese characters by using this tagging

¹<https://hanjaro.juntong.or.kr>

Domains	Train	Validation	Test
Society	67363	2,000	2,000
All	258386	5,000	5,000

Table 3: The statistics for the parallel corpus extracted from Dong-A newspaper (The number of sentences).

information and pre-built dictionary. The UTagger is the Korean morphological tagging model which has a recall of 99.05% on morpheme analysis and 96.76% accuracy on POS and homophone tagging. Nguyen et al. (2019) significantly improved the performance Korean-Vietnamese NMT system by building a lexical semantic network for the special characteristics of Korean, which is using a knowledge base of the UTagger, and applying the Utagger to Korean tokenization.

For MT modeling, we use two types of NMT models: RNN based NMT and Transformer NMT models. We train the NMT models on parallel corpus processed through the Hangul-Hanja conversion above.

5 Experiments

There have been many studies on how to segment Korean and Chinese text (Zhao and Kit, 2008a; Zhao and Kit, 2008b; Zhao et al., 2013; Cai and Zhao, 2016; Deng et al., 2017). To find out which segmentation method has the highest translation performance, we tried multiple segmentation strategies such as byte-pair-encoding (Sennrich et al., 2016), jieba², KoNLP³ and so on. Eventually, we found that character-based segmentation for both languages can give the best performance. Therefore, both Korean and Chinese sentences are segmented into characters for our NMT models.

5.1 Parallel Corpus

We use two parallel corpora in our experiment. The first corpus is a Chinese-Korean parallel corpus of casual conversation and provided by *Semantic Web Research Center*⁴ (SWRC). However, the SWRC corpus contained some incomplete data, so we removed the erroneous data manually. The parallel

corpus consists of a set of 55,294 pairs of parallel sentences. 2,000 and 2,000 pairs from the parallel corpus were extracted as validation data and test data, respectively.

The second corpus (Dong-A) is collected from the online Dong-A newspaper⁵ by us. We collected articles on four domains, Economy (81,278 sentences), Society (71,363), Global (68,073) and Politics (61,208), to build two corpora as shown in Table 3.

Since the sentences in the Dong-A newspaper are relatively long, the maximum sequence length that we used to train the NMT model is set to 200. On the other hand, the maximum sequence length for SWRC corpus is set to 50 because each sentence in the SWRC corpus is short.

5.2 NMT Models

The Torch-based toolkit OpenNMT (Klein et al., 2018) is used to build our NMT models, either RNN-based or Transformer.

As for RNN-based models, we further consider two types of them, one with unidirectional LSTM encoder (uni-RNN) and the other with bidirectional LSTM based encoder (bi-RNN). For both RNN based models, we use 2-layer LSTM with 500 hidden units on both encoder and decoder and use the global attention mechanism as described in (Luong et al., 2015). We use stochastic gradient descent (SGD) optimizer with the initial learning rate 1 and with decay rate 0.5. Mini-batch size is set to 64, and the dropout rate is set to 0.3.

For our Transformer model, both the encoder and decoder are composed of a stack of 6 uniform layers, each built of two sublayers as described in (Vaswani et al., 2017). The dimensionality of all input and output layers is set to 512, and that of Feed-Forward Networks (FFN) layers is set to 2048. We set the source and target tokens per batch to 4096. For optimization, we used Adam optimizer (Kingma and Ba, 2014) with $\beta_1=0.9$, $\beta_2=0.98$ to tune model parameters, and the learning rate is set by the warm-up strategy with steps 8,000, and it decreases proportionally as the model training progresses.

All of the NMT models are trained for 100,000

²<https://pypi.org/project/jieba/>

³<http://konlpy.org>

⁴<http://semanticweb.kaist.ac.kr>

⁵<http://www.donga.com/> (Korean) and <http://chinese.donga.com/> (Chinese)

Systems	BLEU Score (Test set)	
	w/o HH-Conv.	w/ HH-Conv
uni-RNN	33.14	34.44
bi-RNN	35.31	36.66
Transformer	35.47	37.84

Table 4: Experimental results of SWRC corpus. The HH-Conv refers to Hangul-Hanja conversion function.

Systems	Domains	BLEU Score	
		w/o HH-c.	w/ HH-c
uni-RNN	Society	36.25	37.58
	All	39.84	40.70
bi-RNN	Society	39.08	40.00
	All	41.76	42.81
Transformer	Society	39.34	40.55
	All	44.70	44.88

Table 5: Experimental results of Dong-A corpus.

steps and checked the performance on the validation set after every 5,000 training steps. And we save the models every 5,000 training steps and evaluate the models using traditional machine translation evaluation metric.

5.3 Results

We used the BLEU score (Papineni et al., 2002) as our evaluation metric. Tables 4 and 5 show the experimental results for SWRC corpus and Dong-A corpus, respectively. All NMT models, trained with Korean sentences converted through Hangul-Hanja conversion as source sentences, improve the translation performance on all test sets in comparison to the NMT models for the original sentence pairs. The absolute BLEU improvement is about 1.57 on average for SWRC corpus and 0.93 on average for Dong-A corpus when applied the Hangul-Hanja conversion, respectively.

Our proposed method is to improve the translation performance of NMT models by converting only Sino-Korean words into corresponding Chinese characters in Korean sentences using the Hanjaro and sharing the source vocabulary and the target vocabulary.

In the work, we do not convert the entire Korean sentence into Chinese characters using a pre-

specified dictionary and maximum matching mechanism as described in (Zhao et al., 2013). Unlike Chinese, which does not use inflectional morphemes, Korean belongs to an agglutinative language that tends to have a high rate of affixes or morphemes per word. Since some Korean syllables do not have corresponding Chinese characters, so converting all Korean syllables of Korean sentence into Chinese characters is an impossible mission. In fact, we built a bilingual dictionary for Korean and Chinese and used maximum matching mechanism to convert all the affixes and inflectional morphemes of Korean sentences into Chinese characters and trained an RNN based NMT model, but the performance was even lower.

In our implementation, we estimate that the main reason for improving performance is to make the distinction between homophones clearer by converting Sino-Korean words into Chinese characters. Many of the Korean vocabularies that employ the alphabetical writing system are homophones, which can confuse meaning or context. Especially, as mentioned in Section 3, 35% of Sino-Korean words are homophones. Therefore, it is possible to clarify the distinction between homophones by applying Hangul-Hanja conversion to Korean sentences, which leads to performance improvement in Korean-to-Chinese MT.

6 Analysis

6.1 Analysis on Sino-Korean word Conversion

In this subsection, we will analyze the conversion from Sino-Korean words to Chinese characters. To estimate how much Chinese characters converted from Sino-Korean words by Hangul-Hanja conversion function are included in the corresponding reference sentence, we propose *ratio of including the same Chinese character between the converted Korean sentence and Chinese sentence (reference sentence)* (ROIC):

$$ROIC = \frac{\sum_{w_i} f(w_i)}{|w|} \quad (1)$$

where $|w|$ is the number of Chinese words in converted Korean sentence, $f(w_i)$ is 1 if the Chinese word w_i of the converted Korean sentence is included in the corresponding Chinese sentence, and

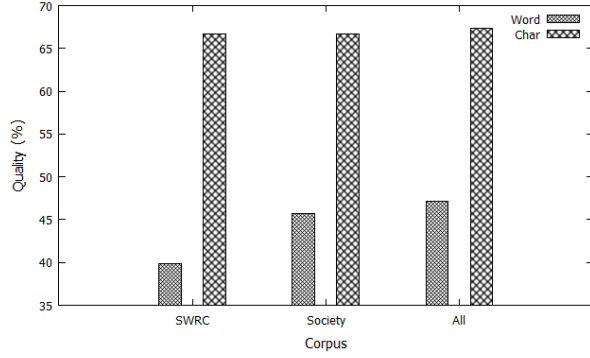


Figure 1: ROIC of each corpus. Word and Char denote the ROIC for Chinese word and the ROIC for Chinese character, respectively.

0 otherwise. For example, in the second example of Table 1, because the five Chinese words such as 两国 (*two countries*), 领域 (*area*), 共同 (*common*), 利益 (*interests*), 确认 (*confirm*) are commonly observed between the converted Korean sentence and the reference sentence except for 广范 (*abroad*), so we say that the ROIC of the converted Korean sentence is $\frac{5}{6}$ (83.33%). We perform analysis of Sino-Korean word conversion in two separate ways: ROIC for Chinese word and ROIC for Chinese character.

Fig. 1 presents the ROIC of each corpus. It can be observed that for each corpus, more than 40% of the converted Chinese words or more than 65% of the converted Chinese characters are included in the reference sentence. So we can see that source vocabulary and target vocabulary share many words after converting Sino-Korean words into Chinese characters. Sharing source vocabulary and target vocabulary is especially useful for same alphabet languages, or for domains where professional terms are written in English (Zhang et al., 2018). Therefore, we set to share the source vocabulary and the target vocabulary of our NMT models, which leads to performance improvement.

6.2 Analysis of Translation Performance according to Different Sentence Lengths

Following Bahdanau et al. (2017), we group sentences of similar lengths together and compute BLEU scores, which are presented in Fig. 2. we conduct this analysis on Society corpus. It shows that our method leads to better translation performance

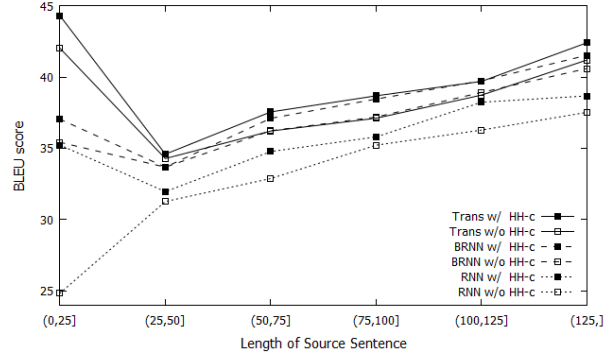


Figure 2: BLEU scores for the translation of sentences with different lengths.

for all the sentence lengths. Since we set the Maximum sentence length to 200 for the Society corpus, we also can see that the performance continues to improve when the length of the input sentence increases.

6.3 Analysis of Homophones Translation

In this subsection, we translate several sentences that contain two homophones and analyze how the Sino-Korean word conversion makes the distinction between homophones more apparent. We translated the sentences using the Transformer model trained with the Dong-A corpus. Table 6 presents the translation results of sentences with two homophones.

We can see that our NMT model clearly distinguishes between homophones for all examples, but the baseline model does not distinguish or translate homophones. For example, in the first example, the baseline model does not translate 유 지* (*community leader*). In the second and third example, the baseline model translated them into the same words without distinguishing between the homophones. In the last example, 의 사** (*wishes*) was improperly translated into 意向 (*intention*). Therefore, as mentioned in Section 5.3, these results indicate that our method helps distinguish homophones in Korean-to-Chinese machine translation.

7 Conclusion

This paper presents a simple novel method exploiting the shared vocabulary of a low-resource language pair for better machine translation. In detail, we convert Sino-Korean words in Korean sen-

Systems	Sentences
Korean	이 지역에 사는 유지*들이 이 마을을 유지**하고 관리해 나가고 있다.
HH-Convert	이 地域에 사는 有志*들이 이 마을을 維持**하고 管理해 나가고 있다.
Chinese	在这个区域生活的有志之士*在维护**和管理这个小区。
English	The <u>community leaders</u> * living in this area are <u>maintaining</u> ** and managing this community.
Trans w/o HH-c	居住在该地区的维持**和管理村庄。
Trans w/ HH-c	居住在该地区的有志*们维持**这个村子，并进行管理。
Korean	이성* 간의 교제는 이성**에 따라 해야 한다.
HH-Convert	异性* 间的 交际는 理性**에 따라 해야 한다.
Chinese	异性*之间交往应该保持理性**。
English	A romantic relationship between the <u>opposite sex</u> * should be <u>rational</u> **.
Trans w/o HH-c	理性**间的交往应遵从理性**。
Trans w/ HH-c	异性*之间的交往应该根据理性**进行。
Korean	그는 천연자원*을 탐사하는 임무에 자원**했다.
HH-Convert	그는 天然资源*을 探查하는 任务에 自願**했다.
Chinese	他自願**参加勘探自然资源**的任务。
English	He <u>volunteered</u> ** for the task of exploring natural <u>resources</u> *.
Trans w/o HH-c	他为探测自然资源**的任务提供了资源**。
Trans w/ HH-c	他自願**担任探测天然资源*的任务。
Korean	의사*의 꿈은 포기했지만, 가족들은 그의 의사**를 존중해주었다.
HH-Convert	医师*의 꿈은 抛弃했지만, 家族들은 그의 意思**를 尊重해주었다.
Chinese	虽然放弃了医生*的梦想,但家人也尊重他的意愿**。
English	Although he gave up on his dream of becoming a <u>doctor</u> *, his family respected his <u>wishes</u> **.
Trans w/o HH-c	虽然医生*的梦想放弃了, 但是家人却尊重了他的意向。
Trans w/ HH-c	虽然放弃了医生*的梦想, 但家人却尊重了他的意愿**。

Table 6: Translation results of sentences with two homophones. The HH-Convert is Korean sentence converted by Hangul-Hanja conversion of the Hanjaro. Trans w/o HH-c and Trans w/ HH-c are the translation results of Transformer baseline model and Transformer using our method, respectively. The underline denotes homophone and the number of stars(*) distinguishes the meanings of the homophone in each example. In Chinese, English, and translation results, they denote words that are equivalent to the homophones in the sense of meaning.

tences into Chinese characters and then train machine translation model with the converted Korean sentences as source sentences. Our proposed improvement has been verified effective over RNN-based and latest Transformer NMT models. Besides, we regard that this is the first attempt which takes a linguistically motivated solution for low-resource translation using NMT models. Although this proposed method seems only suitable for the language pair of Korean and Chinese, it has enormous potential to work for any language pair which shares a considerable vocabulary from their shared history.

References

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv preprint arXiv:1706.03762*.
- Benyamin Ahmadnia, Javier Serrano and Gholamreza Haffari. 2017. Persian-Spanish Low-Resource Statistical Machine Translation Through English as Pivot Language. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017)*, page 24–30.
- Changhyun Kim, Young Kil Kim, Munpyo Hong, Young Ae Seo, Sung Il Yang and Sung-Kwon Choi. 2002. Verb Pattern Based Korean-Chinese Machine Transla-

- tion System. *Proceedings of the First SIGHAN Workshop on Chinese Language Processing*, page 157–165.
- Chenhui Chu, Toshiaki, Nakazawa, Daisuke, Kawahara and Sadao Kurohashi. 2013. Chinese-Japanese Machine Translation Exploiting Chinese Characters. *ACM Transactions on Asian Language Information Processing*, volume 12, page 1–25.
- Chaoyu Guan, Yuhao Cheng and Hai Zhao. 2019. Semantic Role Labeling with Associated Memory Network. *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, page 3361–3371.
- Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu and Feiyue Huang. 2017. Fast and Accurate Neural Word Segmentation for Chinese. *ACL 2017*, page 608–615.
- Deng Cai and Hai Zhao. 2016. Neural Machine Translation of Rare Words with Subword Units. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, page 409–420.
- Dongdong Zhang, Mu Li, Chi-Ho Li and Ming Zhou. 2007. Phrase Reordering Model Integrating Syntactic Knowledge for SMT. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.
- Dong-il Kim, Zheng Cui, Jinji Li and Jong-Hyeok Lee. 2002. A Knowledge Based Approach to Identification of Serial Verb Construction in Chinese-to-Korean Machine Translation System. *Proceedings of the First SIGHAN Workshop on Chinese Language Processing*, volume 18, page 1–7.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.
- Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.
- Fengshun Xiao, Jiangtong Li, Hai Zhao, Rui Wang and Kehai Chen. 2019. Lattice-based Transformer Encoder for Neural Machine Translation. *The 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart and Alexander M. Rush. 2018. OpenNMT: Neural Machine Translation Toolkit. *arXiv preprint arXiv:1805.11462*.
- Hai Zhao and Chunyu Kit. 2008. Exploiting Unlabeled Text with Different Unsupervised Segmentation Criteria for Chinese Word Segmentation. *The 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008)*, volume 33, page 93–104.
- Hai Zhao and Chunyu Kit. 2008. An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework. *The Third International Joint Conference on Natural Language Processing (IJCNLP-2008)*, volume 1, page 9–16.
- Hai Zhao, Masao Utiyama, Eiichiro Sumita, and Bao-Liang Lu. 2013. An Empirical Study on Word Segmentation for Chinese Machine Translation. *CICLing 2013*, page 248–263.
- Hai Zhao, Tianjiao Yin and Jingyi Zhang. 2013. Vietnamese to Chinese Machine Translation via Chinese Character as Pivot. *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, page 250–259.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *In Proceedings of NIPS 2014*, page 3104–3112.
- Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, Hai Zhao, Graham Neubig and Satoshi Nakamura. 2016. Learning local word reorderings for hierarchical phrase-based statistical machine translation. *Machine Translation*, volume 30, page 1–18.
- Jinji Li, Dong-Il Kim and Jong-Hyeok Lee. 2008. Annotation Guidelines for Chinese-Korean Word Alignment. *In Proceedings of LREC*.
- Jinji Li, Jungi Kim, Dong-Il Kim and Jong-Hyeok Lee. 2009. Chinese Syntactic Reordering for Adequate Generation of Korean Verbal Phrases in Chinese-to-Korean SMT. *Proceedings of the Fourth Workshop on Statistical Machine Translation*, page 190–196.
- Jin-Xia Huang and Key-Sun Choi. 2000. Using Bilingual Semantic Information in Chinese-Korean Word Alignment. *Proceedings of the 14th Pacific Asia Conference on Language, Information and Computation*, page 121–130.
- Joon-Choul Shin and Cheol-Young Ock. 2012. A Korean morphological analyzer using a pre-analyzed partial word-phrase dictionary. *KIISE: Software and Applications*, volume 39, page 415–424. [in Korean]
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.
- 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. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, page 1724–1734.

- Lontu Zhang and Mamoru Komachi. 2018. machine translation of logographic language using sub-character level information. *In Proceedings of the Third Conference on Machine Translation*, page 17–25
- Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. *In In Proceedings of the International Workshop on Spoken Language Translation*
- Minh-Thang Luong, Hieu Pham and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, page 1412–1421.
- Nicola Ueffing and Hermann Ney. 2013. Using POS Information for SMT into Morphologically Rich Languages. *10th Conference of the European Chapter of the Association for Computational Linguistics*, page 347–354.
- Nizar Habash and Jun Hu. 2009. Improving Arabic-Chinese Statistical Machine Translation using English as Pivot Language. *Proceedings of the Fourth Workshop on Statistical Machine Translation*, page 173–181.
- Phuoc Nguyen, anh-dung Vo, Joon-Choul Shin, Phuoc Tran and Cheol-Young Ock. 2019. Korean-Vietnamese Neural Machine Translation System With Korean Morphological Analysis and Word Sense Disambiguation. *IEEE Access*, volume 7. page 32602–32614.
- Rafael E. Banchs and Marta Ruiz Costa-jusaa. 2011. A Semantic Feature for Statistical Machine Translation. *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, page 126–134.
- Rico Sennrich and Barry Haddow. 2016. Linguistic Input Features Improve Neural Machine Translation. *Proceedings of the First Conference on Machine Translation*, volume 1, page 83–91.
- Rico Sennrich, Barry Haddow and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.
- Samira Tofighi Zahabi, Somayeh Bakhshaei and Shahram Khadivi. 2013. Using Context Vectors in Improving a Machine Translation System with Bridge Language. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2. page 318–322.
- Shaohui Kuang, Junhui Li, António Branco, Weihua Luo and Deyi Xiong. 2018. Attention Focusing for Neural Machine Translation by Bridging Source and Target Embeddings. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1. page 1767–1776.
- Shuo Zang, Hai Zhao, Chunyang Wu and Rui Wang. 2015. A Novel Word Reordering Method for Statistical Machine Translation. *The 2015 11th International Conference on Natural Computation (ICNC’15) and the 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD’15)*, page 843–848.
- Yingting Wu and Hai Zhao. 2018. Finding Better Subword Segmentation for Neural Machine Translation. *The Seventeenth China National Conference on Computational Linguistics*, Volume 11221, page 53–64.
- Yingting Wu, Hai Zhao and Jia-Jun Tong. 2018. Multilingual Universal Dependency Parsing from Raw Text with Low Resource Language Enhancement. *Proceedings of CoNLL 2018*, page 74–80.
- Yuanmei Lu, Toshiaki Nakazawa and Sadao Kurohashi. 2015. Korean-to-Chinese Word Translation using Chinese Character Knowledge. *Proceedings of MT Summit*, 15(1). page 256–269.
- Zhisong Zhang, Rui Wang, Masao Utiyama, Eiichiro Sumita and Hai Zhao. 2018. Exploring Recombination for Efficient Decoding of Neural Machine Translation. *Proceedings of EMNLP 2018*, page 4785–4790.
- Zuchao Li, Shexia He, Zhuosheng Zhang and Hai Zhao. 2018. Joint Learning for Universal Dependency Parsing. *Proceedings of CoNLL 2018*, page 65–73.
- Zuchao Li, Jiaxun Cai, Shexia He and Hai Zhao. 2018. Seq2seq Dependency Parsing. *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, page 3203–3214.
- Zuchao Li, Shexia He, Hai Zhao, Yiqing Zhang, Zhuosheng Zhang, Xi Zhou and Xiang Zhou. 2019. Dependency or Span, End-to-End Uniform Semantic Role Labeling. *Proceedings of AAAI 2019*.