Research of Uyghur-Chinese Machine Translation System Combination Based on Semantic Information

YaJuan Wang [1,2,3,4], Xiao Li [1,3(✉)], YaTing Yang [1,3(✉)], Azmat Anwar [1,2,3], Rui Dong [1,3]

[1] Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi, China
[4] Department of Information Security Engineering, Xinjiang Police College, Urumqi, China
{ wangyajuan , xiaoli, yangyt , azmat , dongrui }@ms.xjb.ac.cn

Abstract. Uyghur-Chinese Machine Translation System Combination bears some drawbacks of not considering semantic information when doing the combination and the individual systems which participated in system combination lacking diversity. This paper tackles these problems by proposing a system combination method which was generated multiple new systems from a single Statistical Machine Translation（SMT）engine and combined together. These new systems are generated based on a bilingual phrase semantic representation model. Specifically, the Uyghur-Chinese bilingual phrase bilinear semantic similarity score and cosine semantic similarity score were firstly computed by a bilingual phrase semantic representation model and then several new systems were generated by adding features to the original feature set of the phrase-based translation model by static features and dynamic features. Finally, the newly generated system is combined with the baseline system to obtain the final combination results. Experimental results on the Uyghur-Chinese CWMT2013 test sets show that our approach significantly outperforms the baseline by 0.63 BLEU points respectively.

Keywords: Bilingual Phrase Semantic Representation, Statistical Machine Translation, System Combination, Static Feature, Dynamic Feature.

1 Introduction

Although NMT systems have been shown to outperform statistical machine translation (SMT) systems in many rich-resource language pairs translation tasks and for low-resource language pairs translation tasks such as Uyghur-Chinese, the advantage of NMT systems are not obvious due to the limitation of the training data [1,2]. SMT is still one of the most popular methods in Uygur-Chinese machine translation [3]. In recent years, in order to improve the quality of Uyghur-Chinese machine translation, the system combination technology has been applied in Uyghur-Chinese machine translation [4,5]. Although previous work has improved the performance of Uyghur-Chinese machine translation to some extent, many problems remain to be studied. Semantic information has been shown to improve the performance of a single statistical machine translation system[6,7], but whether it will improve the performance of system combination remains to be proven.

The diversity of translation systems is one of the key factors of system combination, so that each combination method cannot be separated from the participation of multiple SMT systems. For Uyghur-Chinese machine translation task, the development of multiple SMT systems requires a considerable cost. From the perspective of effective utilization of resources, it is of great practical significance to derive new systems from a single system and integrate them. So, this paper proposes the method on the Uyghur-Chinese Machine Translation System Combination Based on Semantic Information. It extends the application of semantic information from a single translation system to system combination scenario and uses the extracted semantic similarity as a new feature to generate multiple new systems from a single SMT system and to and to combine multiple systems together.

2 RELATED WORK

With the rapid development of the deep neural network model, more and more researchers have been focused on the semantic representation of bilingual phrases [6-11]. Zhang et al. [6] learned the bilingual phrases representation by using recursive automatic encoder (RAE), and extracted the root node of a binary tree as the semantic representation of bilingual phrases, and proposed a max-margin objective function. Zhang et al. [7] also applied RAE to learn bilingual phrase representation, however they aimed to extract nodes of all levels of binary tree to represent semantic information of different granularity levels (whole phrase, sub-phrase and word) of the whole phrase and propose a two-dimensional attention network to explore the semantic interaction between different levels of granularity.

The most common method of system combination are sentence-level combination and word-level combination [12-17]. However, no matter which combination method is adopted, multiple translation systems are indispensable. A large number of translation systems based on different translation models are not easy to implement for Uyghur-Chinese translation. Therefore, how to achieve system combination without abundantly available systems has attracted much attention. Xiao et al. [18] used bagging and boosting methods to change the distribution of training data and generate a set of "weak translation systems" from the baseline system, and finally used the system combination method to combine these "weak translation systems" to produce a powerful translation system.

Different from their work, we generate new systems by adding semantic features, which are better than the baseline systems due to the introduction of semantic information. Combining these systems with the baseline system can achieve in a more powerful translation result. To make full use of semantic information, a sentence-level fusion method based on N-Best List Reranking is exploited to select a unique hypothesis from multiple translation hypotheses as the final translation result output.

3 System combination Based on Semantic Information

when implementing the Uyghur-Chinese machine translation system combination, the deep neural network model is firstly used to extract the semantic information of Uyghur-Chinese bilingual phrases, and this information is used as a semantic feature to generate a new set of translation systems from a single phrase-based SMT system. Finally, the combination results of the Uyghur-Chinese machine translation system are obtained by combining the generated systems.

3.1 Preparing training data

A key issue in the extraction of bilingual phrase semantic information is how to learn a phrase representation that truly represents the underlying semantics of a bilingual phrase. In order to obtain a large number of translation equivalents (positive phrase pairs) and non-translation pairs (negative phrase pairs) for Uygur-Chinese bilingual phrase, we used forced decoding method [19] method to obtain the positive phrase pair required for the training model. The basic idea of forced decoding is to use the trained decoder to translate the source language data, forcing the decoder to generate the same translation as the reference translation. An example of a positive phrase pairs is shown in Figure 1.

بارىدىغان يول قاردا توسۇلۇپ قېلىپ . ||| 路段 因 暴风雪 阻路 ,
قىزچاق يازغان ماقالە ||| 女孩 写 的 作文
ئاخىرقى 500 مىڭ ||| 最后 五十万
تەشكىللەش رولى ۋە ئاساسى ||| 组织 作用 和 基础
يېڭى يىللىق ئاتكىرتكىنى ||| 新年 贺卡
دەرس مەزمۇنىنى ||| 授课 内容

Fig. 1. examples of positive phrase pairs

Existing completely random substitution method to generate negative phrase pars is not sufficient for the learning of bilingual semantic representation, two different sampling methods are used to generate the negative phrase pairs to enhance the learning ability of the model.

1. Negative phrase pairs are generated by completely random substitute all the words of the Uyghur source phrase and Chinese target phrases.
2. Randomly replace any word in the Uyghur source phrase to generate a Uyghur negative phrase. The Chinese target phrase negative phrase is generated by the same method.

Thus, two negative phrase pairs are generated for each positive phrase pair. Moreover, a positive phrase pair and a negative pair constitutes a training sample of the model. Note that the length of negative phrase pairs should be consistent with that of positive phrase pairs.

3.2 Building bilingual phrase semantic representation model

The goal of learning the semantic representation of bilingual phrases is to measure the semantic similarity between bilingual phrases. However, due to the large difference between Uyghur and Chinese, the clues of the phrase itself are not enough to measure the potential semantic similarity of bilingual phrases. So, the two-dimensional attention-based recursive automatic encoder (RAE) shown in Figure 2 is used as the bilingual phrase semantic representation model.

Firstly, the RAE is used to generate the hierarchical structure of the Uyghur source phrase and the Chinese target phrase respectively. At the same time, multi-level granularity representation is obtained from the generated hierarchical structure.

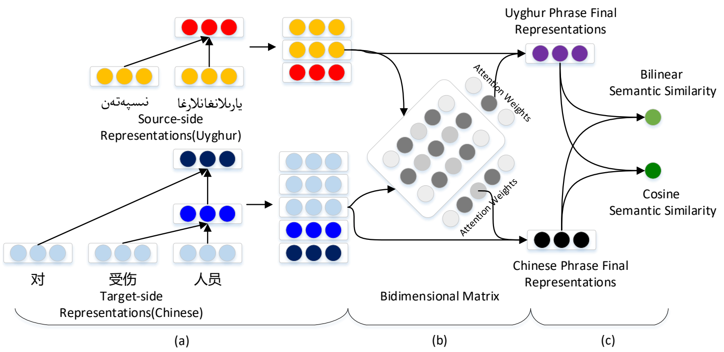

Fig. 2. Uyghur-Chinese bilingual phrase semantic representation model

Secondly, a two-dimensional attention network is constructed here. Based on the learned attention weights, the final representation of the Uyghur source phrase and Chinese target phrase are obtained by integrating the multi-level representation generated from the RAE through equations (2) and (3):

$$p_s = \sum_i a_{s,i} M_{s,i} \tag{1}$$

$$p_t = \sum_j a_{t,j} M_{t,j} \tag{2}$$

Where, $a_{s,i}$ and $a_{t,j}$ are attention weights, $M_{s,i}$ and $M_{t,j}$ indicate the multi-level representation learned by RAE. Finally, after obtaining the semantic representation of the bilingual phrase, the semantic similarity between the Uyghur phrases and Chinese phrases can be calculated. The learned Uyghur phrase representation and Chinese phrase representation need to be mapped to a shared semantic space through the non-linear mapping shown as equation (4) and (5) before calculating semantic similarity.

$$S_s = f\left(W^{(5)} p_s + b^s\right) \tag{3}$$

$$S_t = f\left(W^{(6)} p_s + b^s\right) \tag{4}$$

We evaluate the semantic equivalence between$\mathrm{S}_s$and $\mathrm{S}_t$using bilinear model and cosine model.

1. Bilinear model

Bilinear model compares the source and target phrase representations using a function that is independently linear in both representations. The semantic similarity score is computed as equation (6):

$$s(f,e) = s_s^T S s_t \tag{5}$$

2. Cosine model

Cosine model calculates the cosine value of the angle between $S_s$and S_t . The semantic similarity score is computed as equation (7):

$$s(f,e) = \frac{s_s^T s_t}{\|s_s\|\|s_t\|} \tag{6}$$

3.3 System Combination

generating new system.
We employ the semantic information to extend the original feature space and generate several new systems based on the extension feature space. At present, most of the state-of-art SMT systems are based on linear models. Let $h_m(f,e)$ be a feature function, and λ_m be its weight, a standard SMT model D can be formally written as in equation (8):

$$e^* = \underset{e}{\mathrm{argmax}} \sum_m \lambda_m h_m(f,e) \tag{7}$$

Firstly, we specify Ω to denote the feature space defined by the original set of features used in D . N denotes a new feature space after adding semantic features and $\Omega \subset$ N . According to research by Passban et al. [20], new features can be added in both static and dynamic ways. In this paper, we follow the principle of adding a new semantic feature in the feature space at a time. A new derived system can be generated by the following two methods:

1. Adding static feature

The bilinear similarity score and the cosine similarity score are added behind the original translation model feature h_i to supplement the translation model, to obtain the new features h_i' of the translation model, and to generate new systems in the new feature space N.

2. Dynamic features

The bilinear similarity score and the cosine similarity score are taken as new independent features h_n^b and h_n^c and are added into the linear model to generate new systems in the new feature space N.

After the new generations systems is constructed, each system tunes its feature weights independently to optimize the evaluation metrics on the development set. Let $\mathrm{D} = \{d_1, \cdots, d_n\}$be the set of new systems obtained by adding semantic features. H_i be the n-best list produced by d_i .Then H(D) is the translation candidate list to the system combination model shown in equation (9).

$$\mathrm{H(D)} = \cup_i H_i \tag{8}$$

Constructing of the combination framework.
This paper uses the sentence-level combination method to select the best translation hypothesis from the translation candidate list. This method can also be seen as the translation hypothesis reranking. The method selects the final translation hypothesis according to the scoring function of the equation (10).

$$e^* = \underset{e \in H(D)}{\operatorname{argmax}} \sum_{t=1}^{T} \beta_t \cdot \phi_t(e) + \Psi(e, H_D) \tag{9}$$

where $\phi_t(e)$ is the system score for the t-th single system, β_t is the system weight for the t-th single system that used to represent the preference of the t-th single system. $\Psi(e, H_D)$ is a set of linear combinations based on n-gram features. The framework of the system combination is shown in Figure 3.

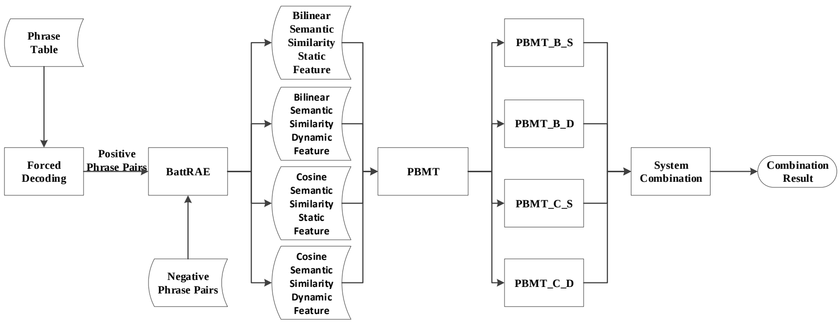

Fig. 3. Uyghur-Chinese machine translation system combination framework

4 Experiments

4.1 Setup

The training corpus used in the experiments is divided into two groups. The first group is the training corpus of CWMT2013, which has about 0.1M parallel sentence pairs. It is used to extract the training data needed for the bilingual phrase semantic representation model. We used forced decoding on the above parallel corpus and collected 1.4M phrase pairs. Each positive phrase pair corresponds to two negative phrase pairs to form two training samples and finally generates about 2.8 M training data for the bilingual phrase semantic representation model. The second group is the training corpus of CWMT2017. After filtering out the sentence pairs repeated with CWMT2013, there are about 0.33M parallel sentence pairs remained, which are used to build a baseline translation system and perform system combination experiments.

For system combination, we building a standard phrase-based Uyghur-Chinese statistical machine translation system on 3.3M parallel corpus of CWMT 2017 using Moses platform as the baseline system. We employ the CWMT2013 development set as the validation data and use the CWMT2013 test set as the test set. We utilized minimum error rate training to optimize the weights of our all translation systems and evaluate the translation quality through the case-insensitive BLEU-4 metric[21]. In order to verify the impact of semantic information on system combination, we set up two experiments for comparison:

1. Semantic information on single system experiment: Add the bilinear similarity score and the cosine similarity score to the baseline system by static features and dynamic features, respectively.
2. Semantic information on system combination experiment: Four new systems are derived based on the added features, and the derived systems are combined with the baseline system.

4.2 Experimental results and analysis

The influence of semantic information on a single system.
We firstly perform an experiment to observe whether the introduction of bilingual phrase semantic information will affect the quality of a single SMT system. By modifying the similarity calculation function of the model, the bilinear similarity score and cosine similarity score are obtained. These two scores are treated as features added to SMT system.

Therefore, according to different similarity scores and different feature application methods, the introduction of semantic features can be divided into bilinear similarity static features (BS), bilinear similarity dynamic features (BD), cosine similarity static features (CS) and cosine similarity dynamics (CD). And when adding two new dynamic features, in order to observe the influence of the initial weight on the translation effect, we set the initial weights from 0.0 to 0.9(interval 0.1), and perform 18 experiments in total. The experimental results on the test set are shown in Figures 4 and 5.

Experimental results show that the initial weight has a direct impact on the translation effect. When adding the bilinear similarity dynamic feature, a good BLEU value is obtained when the initial weight of 0.0; when the cosine similarity dynamic feature is added, the highest BLEU value is obtained when the initial weight is set to 0.1. The experimental results after adding semantic features are shown in Table 1. As shown in Table 1, the introduction of the semantic information of the Uyghur-Chinese bilingual phrases can indeed improve the effect of a single system. The bilinear static feature with the highest score has increased the BLEU value by 0.43 compared with the baseline system. From the similarity method, the bilinear similarity score performs better than the cosine similarity score. From the feature type, the dynamic feature performs better than the static feature, but the introduction of dynamic features requires additional feature functions and different weights need to be set to find the optimal value, so it is more complicated to implement.

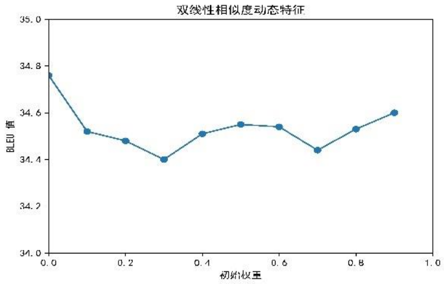

Fig. 4. BLEU value on different weights of bilinear similarity dynamic feature

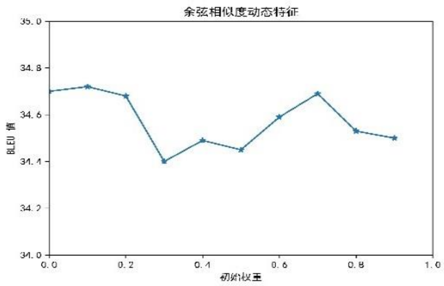

Fig. 5. BLEU value on different weights of cosine similarity dynamic feature

Table 1. Experimental results of semantic features influence on a single system

System	Test	Gains on Test
PBSMT(baseline)	34.38	N/A
PBSMT+B-S(SYS1)	34.81	0.43
PBSMT+B-D(SYS2)	34.76	0.38
PBSMT+C-S(SYS3)	34.55	0.17
PBSMT+C-D(SYS4)	34.72	0.34

The influence of semantic information on system combination.
According to the experimental setup, the system generated from the bilinear similarity static feature is called PBSMT-BS (SYS1) and the system generated from the bilinear similarity dynamic feature is called PBSMT-BD (SYS2). The system generated from the cosine similarity static feature is called PBSMT-CS (SYS3) and the system generated after adding the cosine similarity dynamic feature is called PBSMT-CD (SYS4). In order to observe the impact of the new system on the baseline system, we gradually combined each new system with the baseline system. The results of the system combination are shown in Table 2.

As shown in Table 2, we get the best results of the 35.01 BLEU value after combining the newly generated four systems with the baseline system. Although the BLEU value is decreased after joining the SYS1 system, when with the combination of SYS2 and subsequent SYS3 and SYS4 systems, the effect of system combination is steadily improved compared with the baseline system. The total improvement of the BLEU value is 0.63, which is 0.2 higher than the best single system SYS1. Experimental results show that the semantic information of bilingual phrases can further enhance the effect of system combination on the basis of improving a single translation system.

Table 2. The effect of semantic features on system combination

System	Test	Gains on Test
PBSMT(baseline)	34.38	N/A
+SYS1	34.78	0.4
+SYS2	34.83	0.45
+SYS3	34.86	0.48
+SYS4	35.01	0.63

5 Conclusion

We present a system combination method of Uyghur-Chinese machine translation system based on semantic information, which can introduce semantic information into the process of system combination. Experimental results show that for Uyghur-Chinese translation task, the introduction of bilingual phrase semantic information can not only improve the performance of the single statistical machine translation but also further improve the performance of system combination.

However, the current system combination method is mainly based on statistical methods and due to the fact that the neural machine translation are quite different from statistical machine translation, the existing system combination methods cannot suitable for them. For the future work, we plan to find a new method for system combination which is more suitable for the neural machine translation.

Acknowledgements

This research is supported by the Xinjiang Uygur Autonomous Region Level talent introduction project (Y839031201), National Natural Science Foundation of China (U1703133), Subsidy of the Youth Innovation Promotion Association of the Chinese Academy of Sciences (2017472), the Xinjiang Key Laboratory Fund under Grant (2018D04018)

References

1. Bentivogli, L., Bisazza, A., Cettolo, M., Federico, M.: Neural versus phrase-based machine translation quality: a case study. arXiv preprint arXiv:1608.04631 (2016)
2. Koehn, P., Knowles, R.: Six challenges for neural machine translation. arXiv preprint arXiv:1706.03872 (2017)

3. Li X, Jiang T, Zhou X, Wang L, Yang Y.: Uyghur-Chinese Machine Translation Key Technology Research Overview. Journal of Network New Media 5(1), 19-25 (2016)
4. SU J, Zhang X, Turghun O, Li X.: Joint Multiple Engines Uyghur-Chinese Machine Translation System. Computer Engineering 37(16), 179-181 (2011)
5. Wang Y, Li X, Yang Y, Mi C.: Research of Uyghur-Chinese Machine Translation System Combination Based on Paraphrase Information. Computer Engineering 45(4), 288-295+301 (2019)
6. Zhang, J., Liu, S., Li, M., Zhou, M., Zong, C.: Bilingually-constrained Phrase Embeddings for Machine Translation. pp. 111-121. Association for Computational Linguistics, Baltimore, Maryland (2014)
7. Zhang, B., Xiong, D., Su, J., Qin, Y.: Alignment-Supervised Bidimensional Attention-Based Recursive Autoencoders for Bilingual Phrase Representation. IEEE transactions on cybernetics (2018)
8. Gao, J., He, X., Yih, W., Li, D.: Learning Semantic Representations for the Phrase Translation Model. Computer Science (2013)
9. Su, J., Xiong, D., Zhang, B., Liu, Y., Yao, J., Zhang, M.: Bilingual Correspondence Recursive Autoencoder for Statistical Machine Translation. pp. 1248-1258. Association for Computational Linguistics, Lisbon, Portugal (2015)
10. Zhang, B., Xiong, D., Su, J.: BattRAE: Bidimensional Attention-Based Recursive Autoencoders for Learning Bilingual Phrase Embeddings (2017)
11. Abend, O., Rappoport, A.: The State of the Art in Semantic Representation. pp. 77-89. Association for Computational Linguistics, Vancouver, Canada (2017)
12. Banik, D., Ekbal, A., Bhattacharyya, P., Bhattacharyya, S.: Assembling translations from multi-engine machine translation outputs. Applied Soft Computing 78, 230-239 (2019)
13. Freitag, M., Peter, J.-T., Peitz, S., Feng, M., Ney, H.: Local System Voting Feature for Machine Translation System Combination. pp. 467-476. Association for Computational Linguistics, Lisbon, Portugal (2015)
14. Ma, W.-Y., McKeown, K.: System combination for machine translation through paraphrasing. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1053-1058. (2015)
15. Marie, B., Fujita, A.: 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), pp. 111-124. (2018)
16. Zhu, J., Yang, M., Li, S., Zhao, T.: Sentence-level paraphrasing for machine translation system combination. In: International Conference of Pioneering Computer Scientists, Engineers and Educators, pp. 612-620. Springer, (2016)
17. Freitag, M., Huck, M., Ney, H.: Jane: Open source machine translation system combination. In: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 29-32. (2014)
18. Xiao, T., Zhu, J., Liu, T.: Bagging and boosting statistical machine translation systems. Artificial Intelligence 195, 496-527 (2013)
19. Wuebker, J., Mauser, A., Ney, H.: Training phrase translation models with leaving-one-out. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 475-484. Association for Computational Linguistics, (2010)
20. Passban, P., Hokamp, C., Liu, Q.: Bilingual distributed phrase representation for statistical machine translation. Proceedings of mt summit xv 310-318 (2015)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311-318. Association for Computational Linguistics, (2002)