Sentence-Level Paraphrasing for Machine Translation System Combination

Junguo Zhu, Muyun Yang(✉), Sheng Li, and Tiejun Zhao

Computer Science and Technology, Harbin Institute of Technology,
92 West Dazhi Street, Harbin 150001, China
{jgzhu,ymy}@mtlab.hit.edu.cn, {lisheng,tjzhao}@hit.edu.cn
http://www.hit.edu.cn

Abstract. In this paper, we propose to enhance machine translation system combination (MTSC) with a sentence-level paraphrasing model trained by a neural network. This work extends the number of candidates in MTSC by paraphrasing the whole original MT translation sentences. First we train a neural paraphrasing model of Encoder-Decoder, and leverage the model to paraphrase the MT system outputs to generate synonymous candidates in the semantic space. Then we merge all of them into a single improved translation by a state-of-the-art system combination approach (MEMT) adding some new paraphrasing features. Our experimental results show a significant improvement of 0.28 BLEU points on the WMT2011 test data and 0.41 BLEU points without considering the out-of-vocabulary (OOV) words for the sentence-level paraphrasing model.

Keywords: Machine translation · System combination · Paraphrasing · Neural network

1 Introduction

Machine translation (MT) has made great progress in the past decades of years. Various kinds of MT systems, as represented by Google Translate and Bing Translator, provide diverse translation candidates, which can help people master a rough idea. Each candidate translation is still far from perfect though bearing its own comparative advantage in some condition. Due to the complementary between each other, fusing the translation candidates will be a reasonable solution promoting the quality of MT outputs.

Focusing on integrating multiple MT outputs, many approaches of machine translation system combination (MTSC) have been developed. One of the state-of-the-art solution is a confusion network based method [6,8,9,11,17,20], which splits the candidate sentences into words and reconstruct them as a directed graph. Generating a final translation output is finding an optimized path on the

J. Zhu—This paper is supported by the project of Natural Science Foundation of China (Grant No. 61272384&61370170).

W. Che et al. (Eds.): ICYCSEE 2016, Part I, CCIS 623, pp. 612–620, 2016.
DOI: 10.1007/978-981-10-2053-7_54

graph. Since it implement the combination in word level, the coherence and consistency between words in the original translation is missing. To address the problem, some phrase-level approaches are proposed to add context constrains in combination. Rosti et al. extract source-to-target phrases pairs from the alignments between source sentences and MT system outputs to construct a re-decoding model [19]. Huang and Papineni introduce a hierarchical phrase model is used to handle the source syntax information [10]. A target-to-target phrase-level system combination is construct on word lattice, in which each edge is associated with a phrase (a word or a sequence of words) [4,5,15]. The phrases pairs are extracted form word alignments between a selected best MT candidate and other candidates. One problem of the approach is that the context constrains only prune the branch on existing paths, can not produce the better paths. Ma and McKeown [16] use a hierarchical paraphrasing model to rewrite MT outputs, but they select the most probable one in the rewritten sentences without further combination. In addition, the paraphrases are extracted the alignment between MT outputs, so the approach cannot yet produce new path out of the existing space.

In this paper, a pipeline framework is developed to joins the paraphrasing and translation combination. When paraphrasing, we introduce an encoder-decoder architecture of bidirectional recurrent neural network (RNN) to retain the information of whole sentence. The MT outputs and references are leveraging to train the paraphrasing model, which can extend the space of existing space. As an added benefit, learning the changes from the raw MT outputs to perfect references enable the paraphrasing model to have the ability to correct the translation errors, although it can make new errors. And in the combination process, we introduce a confusion network based combining method with paraphrasing features, which can Our experimental results shows that the effectiveness of our approach is demonstrated on the WMT2011 data of system combination task.

The rest of this paper is organized as follows: Sect. 2 introduces previous studies related to this work. Section 3 presents our method in detail. Section 4 describes and analyzes the experimental results. Finally, Sect. 5 concludes the work with possible future work.

2 Related Works

Focusing on enhancing machine translation system combination, there are two kinds of work related with our method: confusion network based combination and paraphrasing for system combination.

2.1 Confusion Network Based Combination

Bangalore et al. [3] first introduced confusion network into MT system combination using a multiple string alignment algorithm. Confusion network is a weighted directed graph. Each edge is labeled with a word. A translation candidate is represented as a path from the start node to the end node goes through all the other nodes. There are two major directions in current confusion network

research. One is developing monolingual alignment algorithm, such as GIZA++ [17], TER [21], incremental TER [20], ITG [11], IHMM [8], and METEOR [9]. Another one is selecting backbone which is a base of alignment. Sim et al. [21] select the most similar sentence with others according to minimum bayes risk (MBR). Karakos et al. [11] take each sentence as a backbone in one time, and then combine them. Heafield et al. [9] develop a combination system MEMT, which allows the backbone changing in decoder, so the search space is on the top of all the candidates, which is the largest in all previous works.

All above works are based on the existing candidates, extending the number of candidate translations is still an open issue. We aim at further extending the search space to provide more translation candidates by introducing extra paraphrase knowledge. So we use a combination system MEMT [9] as our baseline.

2.2 Paraphrasing for System Combination

Ma and McKeiwn [15] introduce paraphrasing model into combination framework. Mapping the bilingual phrase extraction [14], the paraphrases are extracted according to the word alignments between multiple MT translations. TERp [22] is adopted to obtain word alignments. The phrasal decoder used in the phrase-level combination is based on standard beam search, guided by a log-linear model score. And in their another work, they use a hierarchical paraphrases model to rewrite the MT outputs via CKY algorithm with a Viterbi approximation. Then they combine the combination outputs of the two methods in a sentence-level selection.

Unlike the their works, we rewrite the MT outputs in sentence level. So paraphrasing model can learn the whole context information. And another difference from their works, we extract the paraphrase between MT outputs and references. Since this rewriting is from raw translation to perfect one, it has the ability of correcting errors.

3 Sentence-Level Paraphrasing for System Combination

In this work, we integrate the paraphrasing and system combination in a pipeline style (Fig. 1). Each raw machine translation (MT) candidate is rewritten into a new one. Then we combine MT outputs and rewritten outputs into a final translation.

3.1 Sentence-Level Paraphrasing

Aiming at leveraging the whole sentence information, in our neural paraphrasing work, each translation candidate is represented by a vector $x_1, x_2, ..., x_{T_x}$, also the target translation represented by $y_1, y_2, ..., y_{T_y}$. Similar to a neural machine translation work [1][1], a RNN encoder-decoder architecture is implemented to translate the candidate into target translation.

[1] An open source Toolkit is available at https://github.com/lisa-groundhog/GroundHog.

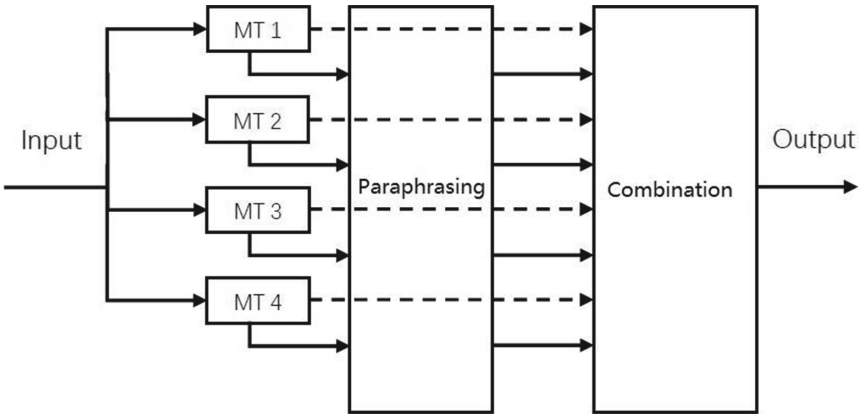

Fig. 1. A Pipeline Framework of Paraphrasing for Translation Fusion

Encoder. In the encoder-decoder framework, a translation candidate $X = (x_1, x_2, ..., x_{T_x})$ is encoded into a fixed-length vector c. $h = (h_1, h_2, ..., h_{T_x})$ by a bi-directional recurrent neural network such that

$$c = q(\{h_1, h_2, ..., h_{T_x}\}) \tag{1}$$

where

$$h_t = \left[\overleftarrow{h_t}; \overrightarrow{h_t}\right] \tag{2}$$

and

$$\overleftarrow{h_t} = f(x_t, \overleftarrow{h}_{t+1}), \overleftarrow{h_t} = f(x_t, \overrightarrow{h}_{t-1}) \tag{3}$$

where h_t is a hidden state at time t, $\overleftarrow{h_t}$ is a forward hidden state and $\overrightarrow{h_t}$ is a backward hidden state. where f and q is some nonlinear functions.

Decoder. In decoder, the context vector c is compute as a weighted sum of these hidden states $h_1, h_2, ..., h_{T_x}$:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{4}$$

The weight α_{ij} of each hidden state is computed as

$$\alpha_{ij} = \frac{\exp(a(h_j, z_{i-1},))}{\sum_{k=1}^{T_x} \exp(a(h_k, z_{i-1}))} \tag{5}$$

$a()$ is a score function of an position alignment model between the inputs and the outputs. z_{i-1} is a RNN hidden state, which emits y_i.

The decoder generates one target word at a time and computes the conditional probability $P(Y|X)$ of translating raw machine translation candidate $x_1, x_2, ..., x_{T_x}$ to target sentence $y_1, y_2, ..., y_{T_y}$ as follows:

$$\log \ P(Y|X) = \sum_{j=1}^{T_y} \log(y_j|y_{<j}, z) \tag{6}$$

where the probability of decoding each word is computed by

$$P(y_j|y_{<j}, z_j) = softmax \ (g(y_{j-1}, z_j, c)) \tag{7}$$

where g is a transformation function that outputs a vocabulary-sized vector.

According to the attention mechanism of this framework, each target word is depended on all the words in translation candidate with different probability α_{ij}. We relieve that all information in translation candidate are encoded into a fixed-length vector.

3.2 System Combination with Paraphrases

To refine the neural paraphrasing outputs into one translation, we connect a translation fusion module in a pipeline style. Since we are not emphasized on the mechanism of system combination itself, here we just chose an existing CN-based combination system MEMT [9] to combine all the MT candidates and paraphrasing candidates. In MEMT framework, the word alignments between difference candidates are implemented by METEOR [2]. The search space is defined on top of all the aligned sentences. A hypothesis starts with the first word of some sentence, and continue to follow the sentence or switch any other sentence. Thus, this pattern can extending the search space as large as possible. A group of features in MEMT are used for scoring partial and complete hypotheses.

- **Length:** the count of words in the hypothesis;
- **LM:** a log probability of language model;
- **Backoff:** average n-gram length found in the language model.
- **Match:** For each n and each system, the number of n-gram matches between the hypothesis and system.

In addition to these original features in MEMT, we also introduce some new features as:

- **Paraphrasing Indicator:** a binary feature in our model to represent if a word is from paraphrasing candidate or not.
- **Paraphrasing Word Counts:** the number of words which from paraphrasing in one sentence.

4 Experiments

4.1 Experimental Settings

In our experiment, we use 2.56 million parallel sentences (machine translation texts and its gold references) to training our neural paraphrasing model. The machine translation texts is provided by a phrase-based machine translation system Moses [13], which is trained on the LDC Chinese-English news corpus[2] In training the RNN encoder-decoder model, the encoder consists of forward and backward recurrent each having 1,000 hidden units, and the decoder also has 1000 hidden units. the word of inputs and outputs is represented by a vector of 620 dimensions. The size of vocabulary in inputs and outputs is set 30,000 words. A minibatch stochastic gradient descent (SGD) algorithm is used to train the neural paraphrasing model. Each SGD update direction is computed using a minibatch of 80 sentences. 350 thousands mini-batches cost 4 days on Taitan X GPU. A beam search is used to find a translation that approximately maximizes the conditional probability [7].

We test our method on the Czech-to-English test data of WMT2011 system combination task. The development set contains 1,003 sentence and test set contains 2,000 sentences. Each sentence has a group of four MT candidates. The translation quality is evaluated by BLEU-4 score [18]. And a statistical significance test is computed by a bootstrap re-sampling method [12] on a 0.1 significant level.

4.2 The Results of Neural Paraphrasing

The BLEU scores of the four MT system on the development and test set are listed in Table 1. And the BLEU score of their rewritten outputs are also listed.

From Table 1, we can find that the paraphrasing model can not improve the raw machine translation for each single MT systems. The greatest cause of these results is that in our methods the whole candidates are completely rewritten by neural paraphrasing model leading to produce some new errors.

But we leverage MEMT to combine 4 way MT translation and 4 way paraphrasing translations, the result is showed in Table 2. For comparison, we also show the highest BLEU score of single system and the BELU score of MEMT.

From Table 2, we can find that our method has an improvement of 1.10 BLEU point than Single best, and an improvement of 0.28 BELU point than MEMT. Clearly, although the neural paraphrasing model can not have an advantage on single MT translations, but in combination process, it can generate helpful candidates with translation knowledge.

However, the size of vocabulary in neural paraphrasing model is a fixed value since a large vocabulary can increase training complexity as well as decoding

[2] LDC2002E18, LDC2002L27, DC2002T01, LDC2003E07, LDC2003E14, LDC2004T07, LDC2005E83, LDC2005T06, LDC2005T10, LDC2005T34, LDC2006E24, LDC2006E26, LDC2006E34, DC2006E86, LDC2006E92, LDC2006E93, LDC2004T08 (HK News, HK Hansards).

Table 1. The Results of Neural Paraphrasing

Systems	Dev	Test
System 1	19.45	18.14
Paraphrasing 1	19.62	18.03
System 2	21.13	20.80
Paraphrasing 2	20.69	20.39
System 3	18.03	17.62
Paraphrasing 3	17.75	17.66
System 4	21.61	21.71
Paraphrasing 4	21.31	21.30

Table 2. The Results of Translation Fusion

	Dev	Test
Single Best	19.45	18.14
MEMT	23.28	22.74
MEMT+Paraphrasing	23.60	23.02

Table 3. The Results of Translation Fusion without OOV

	Test
Single Best	21.18
MEMT	21.87
MEMT+Paraphrasing	22.28

complexity. Hence, the OOVs will affect the final translation performance. We filter those sentence with OOV words from the test set. We list the results Single best, MEMT, and MEMT+Paraphrasing, on test set without OOVs in Table 3.

In Table 3, we present that our method has a further improvement of 0.41 BLEU points. Obviously, neural paraphrasing is able to expanding the decoding space for translation fusion by generating new candidates, while our combination model is able to grasp the useful information by consistency decoding.

5 Conclusion

In this work, we propose a new pipeline framework to integrate the paraphrasing and MT system combination. We train a RNN encoder-decoder paraphrasing model to provide more translation candidates for system combination. Our neural paraphrasing model is trained on the whole sentence, which can retain sentence-level information. And we also use a state-of-the-art combination system with some paraphrasing features to combine the original translation and

their rewritten translations. The experimental results show an effective performance at an improvement of 0.28 BLEU point on the WMT11 test and 0.41 on the test set without OOVs compared with MEMT.

Our contributions are as following: (1) We extending the number of candidates for system combination via paraphrasing to extend the search space of the final translations. (2) We introduce a neural network to exploit the sentence-level information. And (3) we use the MT outputs and references to train the paraphrasing model, which can correct the errors in MT outputs.

In this work, we find that the RNN-based paraphrase model can not improve the translation directly, so further work will try more neural network frameworks to address the task of MT system combination.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014). arXiv:1409.0473
2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 (2005)
3. Bangalore, S., Bordel, G., Riccardi, G.: Computing consensus translation from multiple machine translation systems. In: IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2001, pp. 351–354. IEEE (2001)
4. Du, J., Way, A.: Using terp to augment the system combination for SMT. In: Proceedings of the Ninth Conference of the Association for Machine Translation. Association for Machine Translation in the Americas (2010)
5. Feng, Y., Liu, Y., Mi, H., Liu, Q., Lü, Y.: Lattice-based system combination for statistical machine translation. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 3, pp. 1105–1113. Association for Computational Linguistics (2009)
6. Freitag, M., Peter, J.T., Peitz, S., Feng, M., Ney, H.: Local system voting feature for machine translation system combination. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, pp. 467–476 (2015)
7. Graves, A.: Sequence transduction with recurrent neural networks (2012). arXiv:1211.3711
8. He, X., Yang, M., Gao, J., Nguyen, P., Moore, R.: Indirect-HMM-based hypothesis alignment for combining outputs from machine translation systems. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 98–107. Association for Computational Linguistics (2008)
9. Heafield, K., Lavie, A.: Combining machine translation output with open source: the Carnegie Mellon multi-engine machine translation scheme. Prague Bull. Math. Linguist. **93**, 27–36 (2010)
10. Huang, F., Papineni, K.: Hierarchical system combination for machine translation. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 277–286. Association for Computational Linguistics, Prague, Czech Republic, June 2007

11. Karakos, D., Eisner, J., Khudanpur, S., Dreyer, M.: Machine translation system combination using ITG-based alignments. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, pp. 81–84. Association for Computational Linguistics (2008)
12. Koehn, P.: Statistical significance tests for machine translation evaluation. In: EMNLP, pp. 388–395. Citeseer (2004)
13. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 177–180. Association for Computational Linguistics (2007)
14. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1, pp. 48–54. Association for Computational Linguistics (2003)
15. Ma, W.Y., McKeown, K.: Phrase-level system combination for machine translation based on target-to-target decoding. In: Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas (2012)
16. 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)
17. Matusov, E., Ueffing, N., Ney, H.: Computing consensus translation for multiple machine translation systems using enhanced hypothesis alignment. In: EACL, pp. 33–40 (2006)
18. 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)
19. Rosti, A.v.I., Ayan, N.F., Xiang, B., Matsoukas, S., Schwartz, R., Dorr, B.J.: Combining outputs from multiple machine translation systems. In: Proceeding NAACL-HLT 2007, pp. 228–235 (2007)
20. Rosti, A.V.I., Matsoukas, S., Schwartz, R.: Improved word-level system combination for machine translation. In: Annual Meeting-Association for Computational Linguistics, pp. 312–319 (2007)
21. Sim, K.C., Byrne, W.J., Gales, M.J.F., Sahbi, H., Woodland, P.C.: Consensus network decoding for statistical machine translation system combination. In: IEEE International Conference on Acoustics Speech Signal Processing, ICASSP 2007, vol. 4, pp. 2–5 (2007)
22. Snover, M.G., Madnani, N., Dorr, B., Schwartz, R.: TER-Plus: paraphrase, semantic, and alignment enhancements to translation edit rate. Mach. Transl. **23**(2–3), 117–127 (2009)