IEICE TRANS. INF. & SYST., VOL.E97–D, NO.1 JANUARY 2014

LETTER

Sentence-Level Combination of Machine Translation Outputs with Syntactically Hybridized Translations

Bo WANG[†a)], ***Member***, **Yuanyuan ZHANG**[†], ***and*** **Qian XU**[†], ***Nonmembers***

SUMMARY We describe a novel idea to improve machine translation by combining multiple candidate translations and extra translations. Without manual work, extra translations can be generated by identifying and hybridizing the syntactic equivalents in candidate translations. Candidate and extra translations are then combined on sentence level for better general translation performance.

key words: *combination, machine translation, automatic extension, syntactic equivalent*

1. Introduction

System combination has been widely explored in machine translation (MT) systems, which can be done on sentence-level, phrase-level or word-level [1]–[4].

Sentence-level and word-level combination are two most popular strategies which has strength and weakness in different aspects, respectively.

Sentence-level methods are based on the re-ranking of a merged N-best hypothesis list of the MT systems. A confidence score is assigned to each hypothesis and the most confident one is selected as the best translation. The process is more complex for word-level combination. Firstly, hypotheses in the N-best list are decomposed into words. And then the words are re-decoded to construct an optimized new hypothesis. In theory, word-level combination is more effective, while sentence-level combination is more robust.

Word-level combination can generate new hypotheses constructed with the well translated words from different MT systems. Sentence-level combination only selects among the existing translations containing both the well translated and badly translated words of a single system.

Sentence-level combination is more faithful to original translation and thus assures less risk. Well translated phrases serve as natural units semantically. Therefore, in word-level combination, breaking a coherent phrase is a risk: separated words are possibly to be re-organized into worse phrases. Word-level combination also suffers the mistakes of alignment, especially for the MT systems having different architecture.

In general, the performance of sentence-level combination highly depends on the quality and quantity of the candidate translations. In an extreme case, if the candidate translations involve all possible alternations of translations, the sentence-level combination could be much more effective than word-level combination. Consequently, the sentence-level combination can be improved by increasing the number of translation alternations, and retain the advantage of robustness.

To obtain more translations, if we force the candidate MT systems to generate more translations, the quality of extra translations will be much poorer. On the contrary, in this work, we examine the existing high-quality candidate translations, and find the phrases which can be exchanged with the syntactic information. In this way, we generate extra translations of relatively high quality automatically, and thus improve the sentence-level combination by improving the translation coverage.

2. Extension of the Candidate Translations

2.1 Syntactic Equivalents

There are two kinds of variations of the sentences having the same meaning. One is structural variation, in which presentations employ the same words in different structure. The other is lexical variation, in which presentations employ the different words in same structure. In practice, one candidate translation sentence often has both of the two kinds of variations comparing with other candidate translation sentences.

We focus on the lexical variations. We firstly propose a method to identify equivalent sub-segments among translations with syntactic information. A sub-segment is a words sequence of arbitrary length. In our work, the equivalents of a sub-segment s in a candidate translation sentence are identified as the sub-segments playing the same syntactic role [5]–[7] in corresponding candidate translations. The equivalents obtained in this way are called syntactic equivalents.

Suppose S_1 and S_2 is a sentence pair sharing the same meaning. T_1 and T_2 are the consecutive syntactic trees of S_1 and S_2 respectively. A syntactic equivalent pair can be formally defined between S_1 and S_2 with a 4-tuple: $< n_1, n_2, s_1, s_2 >$, where n_i is a non-terminal node in T_i and s_i is the sub-segment which is covered by n_i. Then, all the syntactic equivalent pair S_1 and S_2 can be recursively identified using following process:

- The first syntactic equivalent pair $< n_1, n_2, s_1, s_2 >$ is identified where n_i is the root of T_i and $s_i = S_i$.

Manuscript received July 9, 2013.
Manuscript revised August 23, 2013.

[†]The authors are with School of Computer Science and Technology, Tianjin University, China.
a) E-mail: bo.wang.1979@gmail.com

DOI: 10.1587/transinf.E97.D.164

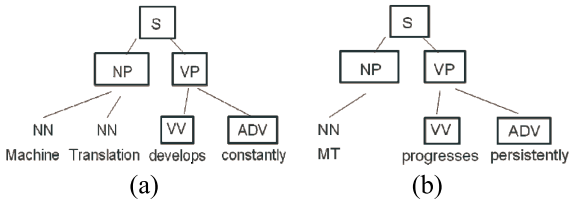

Fig. 1 An example of the syntactic equivalents.

- Suppose $< n_1, n_2, s_1, s_2 >$ is a syntactic equivalent pair. $\{n_{11}, n_{12}, \ldots, n_{1m}\}$ and $\{n_{21}, n_{22}, \ldots, n_{2n}\}$ are the child nodes sequences of n_1 and n_2 respectively. If $n = m$ and $n_{1i} = n_{2i}$ (i.e. the child nodes sequence of n_1 and n_2 are exactly the same), for each node pair n_{1i} and n_{2i} a syntactic equivalent pair is identified as $< n_{1i}, n_{2i}, s_{1i}, s_{2i} >$.

Figure 1 is an example of the syntactic equivalents. The nodes which are identified as syntactic equivalents are tagged by a rectangle.

In this example, five equivalent pairs can be identified:

- <S, S, "Machine translation develops constantly", "MT progresses persistently">
- <NP, NP, "Machine translation", "MT">
- <VP, VP, "develops constantly", "progresses persistently">
- <VV, VV, "develops", "progresses">
- <ADV, ADV, "constantly", "persistently">

2.2 Hybridization of Syntactic Equivalents

The syntactic equivalents share the same role in the same syntactic structure. Therefore, we can obtain the variation of two translations by exchange the sub-segments of syntactic equivalents between translations. Furthermore, multiple extra translations can be generated by exchanging syntactic equivalents among multiple translations. This is called the syntactic hybridization which can be illustrated by following steps:

Suppose $S = \{S_i\}_{i=0\ldots n}$ is a sentence set containing n candidate translation sentences. S' is the new sentence set containing the original sentences and extra hybridized sentences. S' can be obtained by Eq. (1):

$$S' = \bigcup_{i=0}^{n} Equ(root_i) \tag{1}$$

where $root_i$ is the root of the syntactic tree of S_i. $Equ(nt)$ returns the set of all equivalent of the sub-segments covered by the tree node nt, i.e., $Equ(root_i)$ returns all equivalent variations of S_i. The detailed process of $Equ(nt)$ is:

***Equ*(*nt*) :**
Define set equ = Φ
Add Seg(*nt*) *to equ*
If nt is included in an equivalent pair $< nt, nt', s, s' >$
    Add s' *to equ*
    Define $child_{i=1\ldots m}$ *are the m children of nt*
    Define hybr = $Equ(child_1) \times Equ(child_2) \ldots \times Equ(child_m)$
    Merge hybr into equ
Return equ

where $Seg(nt)$ is the sub-segment covered by the nt. Operation $Equ_1 \times Equ_2$ generates the Cartesian product of the sub-segment sets Equ_1 and Equ_2, i.e., for each sub-segment pair s_1 and s_2 from Equ_1 and Equ_2 respectively, we concatenate s_1 and s_2.

For the example in Fig. 1, six hybridized translations can be generated besides the original two sentences:

- Machine Translation develops persistently
- Machine Translation progresses constantly
- Machine Translation progresses persistently
- MT develops constantly
- MT develops persistently
- MT progresses constantly

3. Sentence-Level Combination with Weighted MBR

Minimum Bayes-Risk (MBR) decoding is proposed in Ref. [8] finding a hypothesis with the lowest Bayes risk with respect to all the candidate translations. In previous work, we propose wMBR to improve MBR fitting the MBR to certain tasks. wMBR weights the hypotheses with the general performance of the MT systems instead of the sensitive features on sentence or word level. wMBR is supposed to be a balance between the under-fitting and over-fitting. Equation (2) formulates the principle of wMBR.

$$E_{wMBR} = \arg\min_{E'} \frac{1}{pf_{E'}} \sum_{E} pf_E P(E|F) L(E, E') \tag{2}$$

where Pf_E is the general performance score of the MT system which generates the hypnosis E. $P(E|F)$ is the posterior probability of the hypnosis E. $L(E, E')$ is the expected loss of E and E' calculated as the reciprocal of the similarity between E and E'.

In this work, we firstly extend the candidate translations by syntactic equivalents hybridization. Then, the wMBR sentence-level combination is performed on the extended set, and selects the best translation.

4. Experiments

4.1 Experimental Settings

The corpus for the experiments is from the combination track of CWMT (China workshop of Machine Translation). In CWMT, the participants of machine translation track are asked to provide the translations of the development and test corpus of the MT track with their candidate MT systems. Then, the translations of the development corpus and test corpus are provided as the training corpus and test corpus for

the combination track respectively. In totally, outputs from 17 candidate MT systems are available for the combination track. It's noticeable that, candidate MT systems whose performance is much worse than the best candidate system often have negative effects on the combination. Therefore, in our experiments, for both development and test, only the outputs from the MT system which achieve the top 5 performance on the development corpus were selected to be combined. Another reason to select the best translations is to obtain better parsing results.

In combination, we ask for the 10-best translations for each source sentence and each input MT system. Each translation is parsed by the Stanford parser to identify the syntactic equivalents and generate the extra translations. Both original sentences and extra sentences are merged into one set for combination.

The typical MBR and wMBR combination methods are used as baselines in our experiments:

- **wMBR-BLEU / NIST / TER / GTM**: the wMBR-Based methods using BLEU [9] / NIST [10] / TER [11] / GTM [12] for the measuring of the general performance of the input MT systems and the similarity between the hypotheses.
- **MBR-BLEU**: the MBR-Based method using BLEU for the similarity measuring.
- **Multi-Features:** an implementation of the method in Ref. [2]. This method adopts a generalized linear models and a set of sophisticated sentence-level features to obtain a confidence for each unique hypothesis and get new ranking.

4.2 Experimental Results

Table 1 lists the BLEU score for each combination system and the sentence-level upper bound (sentence-level upper bound is calculated by selecting the candidate translation which matches the reference best) on the development corpus. Table 2 lists the same scores on the test corpus. In two tables, the results on original outputs set ("Original") and the results on automatic extended outputs set ("+Extension") are both included. The paired t-statistic scores ("t-score") of each pair of BLEU scores are also calculated following [13].

As shown in the results, in most cases, the automatically extended outputs sets can improve the performance of the combination methods significantly (> 95%, i.e., t-score > 2.262).

It's noticeable that on both development and test data, the automatic extensions of the outputs sets increase the upper bound of the combination. This is very important because the improvement of the upper bound is a fundamental improvement which is independent of the combination methods. This kind of improvement increases the space for the methods to be improved.

Table 1 Results of sentence-level combinations on development data.

Combination Methods	*Original*	*+Extension*	*t-score*
wMBR-BLEU	0.3420	0.3422	1.50
wMBR-NIST	0.3392	0.3389	-1.20
wMBR-TER	0.3269	0.3275	3.60
wMBR-GTM	0.3376	0.3510	14.78
MBR-BLEU	0.3351	0.3411	4.61
Con_wMBR -BLEU	0.3370	0.3370	0.75
Multi-Features	0.3402	0.3566	13.89
Sen-Level Upper Bound	**0.4102**	**0.4166**	14.31

Table 2 Results of sentence-level combinations on test data.

Combination Methods	*Original*	*+Extension*	*t-score*
wMBR-BLEU	0.2944	0.2948	3.00
wMBR-NIST	0.2907	0.2978	15.43
wMBR-TER	0.2590	0.2611	10.50
wMBR-GTM	0.2771	0.2771	-1.00
MBR-BLEU	0.2793	0.2787	-0.48
Con_wMBR -BLEU	0.2808	0.2810	2.40
Multi-Features	0.2192	0.2190	-1.50
Sen-Level Upper Bound	**0.3347**	**0.3412**	17.73

Table 3 Counts of the tree nodes and equivalent nodes in translations.

Average words count of candidate translations sentence	36.62
Total tree nodes	2835576
Total equivalent nodes	362349
Average tree nodes per sentence	56.37
Average equivalent nodes per sentence	7.20

4.3 Distribution of the Equivalents

To investigate the distribution of the equivalents, we perform several statistics about the count and the length of the syntactic nodes in test data. Table 3 lists the information about the count of the nodes. The first row is the average words count per candidate translation sentence. The second and third row is the count of all tree nodes and equivalent nodes in all candidate translations sentence, respectively. The fourth and fifth row is the average count of tree nodes and equivalent nodes per candidate translations sentence respectively. We can see the ratio of the syntactic equivalent nodes to all syntactic nodes is 12.8%.

We also investigate the distribution of the length (count of covered words) of the nodes. First, we count the tree nodes and equivalent nodes whose length is from 1 word to 50 words. Then we calculate the proportion between equivalent nodes and tree nodes for each length. Figure 2 illustrates the distribution of absolute count of the equivalent nodes. The X-axis is the length of the nodes and the Y-axis is the count. Figure 3 illustrates the distribution of the proportions between equivalent nodes and tree nodes in same length. The X-axis is the length of the nodes and the Y-axis is the proportion.

The investigation reveals three messages. First, the absolute counts of the short equivalents are much greater than those of long equivalents. Second, the proportion of the equivalents of middle length is greater than those of short equivalents, this clarify that the reason of large amount of

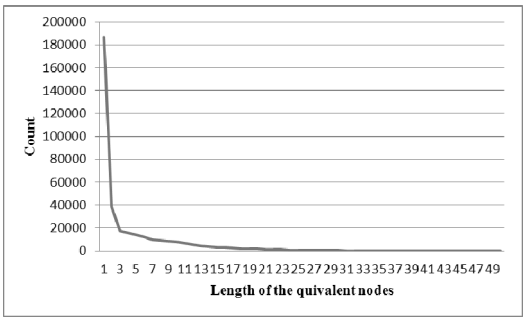

Fig. 2 Distribution of equivalent node of different lengths.

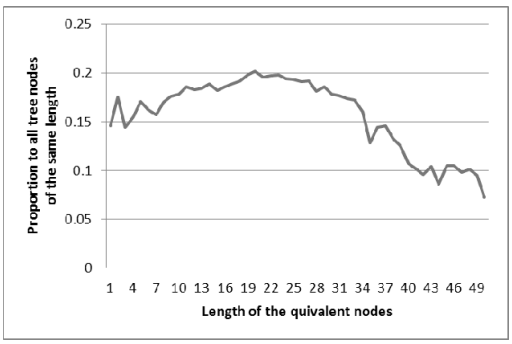

Fig. 3 Distribution of proportion of equivalent nodes to all tree nodes of the same length.

short equivalents is the large amount of short tree nodes. Third, we can see that the new method comparably bias to the longer equivalents. This happens because the method adopts a top-down survey of the tree.

4.4 Compared with Word-level Combination

The proposed hybridization of syntactic equivalents can be used as preprocess of both sentence-level and word-level combination, but, in theory, it can hardly help the word-level combination which also have the ability of reorganizing the existing words. On word-level, Watanabe [3] proposed to guide the confusion of word network with syntactic structure, and prove that syntactic information is also helpful to the word-level combination.

Another consideration is that is the "hybridization preprocess + sentence-level combination" better than word-level combination? Though the word-level combination is more advanced in theory, sentence-level method still has several advantages: relieving or omitting the training process, more stable across heterogeneous datasets and less risk in breaking coherent phrases. Well trained word-level combination has been proved to be advanced on certain datasets, while sentence-level methods may be more robust in open tasks across heterogeneous datasets, which should be explored with large-scale experiments in our future study.

5. Conclusion

In this paper, we propose an idea to improve the sentence-level combination of the outputs of MT systems. Instead of a new combination process, we introduce a novel pre-process to automatically increase the coverage of candidate translation without more MT systems or forcing MT systems to generate more translations. In pre-process, the outputs set is extended by the identification and hybridization of the syntactic equivalents among existing translations, and then extra translations are generated by exchanging the equivalents.

The experimental results indicate that the additional pre-process can not only improve the performance of various sentence-level combination systems, but also increase the upper bound of the sentence-level combination which is a fundamental improvement.

Acknowledgments

This work is supported by the Chinese National Program on Key Basic Research Project (973 Program, grant no.2013CB329304), the Natural Science Foundation of China (No.61105072 and 61272265) and Tianjin Key Laboratory of Cognitive Computing and Application.

References

[1] F. Huang and K. Papineni, "Hierarchical system combination for machine translation," Proc. EMNLP, 2007.
[2] A. Rosti, et al., "Improved word-level system combination for machine translation," Proc. 45th ACL, 2007.
[3] T. Watanabe and E. Sumita, "Machine translation system combination by confusion forest," Proc. HLT, 2011.
[4] K. Sim, et al., "Consensus network decoding for statistical machine translation system combination," Proc. IEEE Conference on ASSP, 2007.
[5] B. Wang, et al., "References extension for the automatic evaluation of MT by syntactic hybridization," SSST3 Workshop, NAACL HLT, Boulder, Colorado, June 2009.
[6] B. Prachya and S. Thepchai, "Probabilistic treatment for syntactic gaps in analytic language parsing," IEICE Trans. Inf. & Syst., vol.E94-D, no.3, pp.440–447, March 2011.
[7] Y. Kato and S. Matsubara, "Incremental parsing with adjoining operation," IEICE Trans. Inf. & Syst., vol.E92-D, no.12, pp.2306–2312, Dec. 2009.
[8] R. Tromble, et al., "Lattice minimum bayes-risk decoding for statistical machine translation," Proc. EMNLP, 2008.
[9] K. Papieni, et al., "BLEU: A method for automatic evaluation of machine translation," Proc. ACL, 2002.
[10] G. Doddington, "Automatic evaluation of MT quality using N-gram co-occurrence statistics," Proc. HLT, 2002.
[11] M. Snover, et al., "A study of translation edit rate with targeted human annotation," Proc. AMTA, 2006.
[12] I.D. Melamed, et al., "Precision and recall of machine translation," Proc. HLT/NAACL, 2003.
[13] P. Koehn, "Statistical significance tests for machine translation evaluation," Proc. EMNLP, 2004.