References Extension for the Automatic Evaluation of MT by Syntactic Hybridization

Bo Wang, Tiejun Zhao, Muyun Yang, Sheng Li
School of Computer Science and Technology
Harbin Institute of Technology
Harbin, China
{bowang,tjzhao,ymy,sl}@mtlab.hit.edu.cn

Abstract

Because of the variations of the languages, the coverage of the references is very important to the reference based automatic evaluation of machine translation systems. We propose a method to extend the reference set of the automatic evaluation only based on multiple manual references and their syntactic structures. In our approach, the syntactic equivalents in the reference sentences are identified and hybridized to generate new references. The new method need no external knowledge and can obtain the equivalents of long subsegments of reference sentences. The experimental results show that using the extended reference set the popular automatic evaluation metrics achieve better correlations with the human assessments.

1 Introduction

While human evaluation of machine translation output remains the most reliable method to assess translation quality, it is a costly and time consuming process. The development of automatic machine translation evaluation metrics enables the rapid assessment of system output. By providing immediate feedback on the effectiveness of various techniques, these metrics have guided machine translation research and have facilitated rapid advances in the state of the art. In addition, automatic evaluation metrics are useful in comparing the performance of multiple machine translation systems on a given translation task. Since automatic evaluation metrics are meant to serve as a surrogate for human judgments, their quality is determined by how well they correlate with assessors' preferences and how accurately they predicts human judgments.

Although current methods for automatically evaluating machine translation output do not require humans to assess individual system output, humans are nevertheless needed to generate a number of reference translations. The quality of machine-generated translations is determined by automatically comparing system output with these references. All current automatic evaluation metrics are based on the various measures of the general similarity between the system translation and manual references. This kind of method has an obvious drawback: it does not account for combinations of lexical and syntactic differences that might occur between a perfectly fluent and accurately-translated machine output and a human reference translation (beyond variations already captured by the different reference translations themselves). Moreover, the set of human reference translations is unlikely to be an exhaustive inventory of "good translations" for any given foreign language sentence. Therefore, it would be highly desirable to extend the coverage of the references for the similarity based evaluation methods.

To match the system translation with various presentation of the same meaning, many work haven been proposed to extend the references by generating lexical variations. The first strategy focuses on the extension based on paraphrase identi-

Proceedings of SSST-3, Third Workshop on Syntax and Structure in Statistical Translation, pages 37–44,
Boulder, Colorado, June 2009.

fication (Lepage and Denoual, 2005; Lassner et al. 2005; Zhou et al. 2006; Kauchak and Barzilay, 2006; Owczarzak et al. 2006; Owczarzak et al. 2007). In this kind of method, the quality of system translations can be viewed as the extent to which the conveyed meaning matches the semantics of the reference translations, independent of substrings they may share. In short, all paraphrases of human-generated references should be considered "good" translations. The second strategy extends the references with the synonymy (Banerjee and Lavie, 2005; Lassner et al. 2005). This is an alternation to obtain lexical variations with synonymy dictionaries instead of the paraphrase. In this kind of method, the reference is matched against to the system translation with the pack of the synonymies of the reference words instead of the exact matching.

Both two strategies can successfully capture the lexical variations and greatly extend the coverage of the references. But they still have two common deficiencies. The first is the demand of the external knowledge. Paraphrase based method need a mass of external corpus to extract paraphrases and synonymy based method need manually constructed semantic dictionaries. These demands seriously limit the application on various languages for which the external knowledge is absent.

Another deficiency is that the two strategies cannot capture the equivalents of long subsegments such as a clause. Synonymy based method can only capture the equivalents of single words. Paraphrase based method can capture the equivalents of longer units but the length is still very narrow. In many cases, some long subsegments can be varied with an entirely different presentation which cannot be decomposed into the variations of words or phrases.

To address these problems we propose a novel strategy to generate variations presentation only using existing multiple manual references without any external knowledge. We identify the syntactic components on different level as the replaceable units and determine the syntactic equivalents of the components in the corresponding references. Then the equivalents of the syntactic components are hybridized into new references.

The rest of the paper is organized as follows. Section 2 introduces the concept and identification of the syntactic equivalents. Section 3 proposes a process to hybridize the syntactic equivalents efficiently. Experimental results are illustrated in section 4. We also include some related discussion in Section 5. Finally this work is concluded in Section 6.

2 Syntactic Equivalents

In our approach, we propose a novel method to obtain the equivalents of the sub-segments from the corresponding references to a single source sentence. A sub-segment can be a word, a phrase or longer unit such as a clause. As we know, the variations of the sentences to the same meaning can be distinguished into two categories. The first is the structural variations. In this case, presentations employ the same words but arrange them in different structure. The second is lexical variations. In this case, presentations have the same structure but employ the different words. In practice, one reference sentence often has both of the two kinds of variations comparing with other corresponding reference sentences.

As the previous works, we also focus on the lexical variations. The approach is that the equivalents of the words are not obtained by external knowledge. In our strategy, generally speaking, the equivalents of a sub-segment S in a reference sentence are identified as the sub-segments which play the same syntactic role in the same structure in the other corresponding references. The equivalents obtained in this way are called syntactic equivalents.

Suppose R_1 and R_2 is a corresponding reference sentence pair. T_1 and T_2 are the consecutive syntactic trees of R_1 and R_2 respectively. We formally define a syntactic equivalent pair between R_1 and R_2 with a 4-tuple:

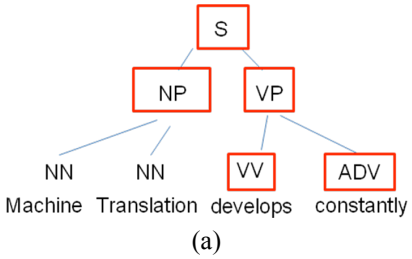
$$<N_1, N_2, S_1, S_2>$$

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 R_1 and R_2 can be recursively identified using following process:

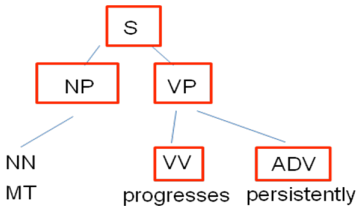
- The first syntactic equivalent pair $<N_1, N_2, S_1, S_2>$ is identified where N_i is the root of T_i and $S_i = R_i$.
- Suppose $<N_1, N_2, S_1, S_2>$ is a syntactic equivalent pair. $\{N_{11}, N_{12}, \ldots N_{1m}\}$ and $\{N_{21}, N_{22}, \ldots N_{2n}\}$ are the child nodes sequences of

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 $< N_{1i}, N_{2i}, S_{1i}, S_{2i}>$.

With this process, all equivalent pairs on different syntactic level can be identified by synchronously traveling the two trees from top to bottom. The following is an example of the identification of the equivalent pairs. Figure 1 gives out a reference sentence pair and their syntactic trees. The nodes which are included in certain equivalent pair are surrounded by a rectangle.



(a)



(b)

Figure 1 An example of the identification of the syntactic equivalent pairs.

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">*

3 Hybridization of Syntactic Equivalents

The indentified syntactic equivalents pairs include the sub-segments which sharing the same role in the same syntactic structure. Because of this, we can obtain a variation of a reference sentence by switching the two sub-segments of an equivalent pair in this sentence. This operation did not change the structure of the sentence but only replace a sub-segment in the structure with its equivalent.

Consequently, two new references can be generated by switching the two sub-segments of an equivalent pair between two reference sentences. Furthermore when we switch the sub-segments of all equivalent pairs between the two references, multiple new references are generated with various combinations of the switches. This operation is called the syntactic hybridization of the references which can be illustrated by following steps:

Suppose $R=\{r_i\}_{i=1...n}$ is a reference set containing n reference sentences to a single source sentence. *R'* is the new reference set containing the original reference sentences and the hybridized reference sentences. *R'* can be obtained by formula (1):

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

where $root_i$ is the root node of the syntactic tree of r_i. *Equ(nt)* returns the set of all equivalent of the sub-segments covered by the tree node *nt*. 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 p' to equ
Define $child_{i=1...m}$ *is 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 tree node *nt*. Operation $S_1 \times S_2$ generates the Cartesian product of the sub-segment set S_1 and S_2, i.e. for each arbitrary sub-segment pair s_1 and s_2 selected from S_1 and S respectively, we concatenate s_1 and s_2. Finally, the reduplicate references in *R'* are removed.

For the example in Section 2, eight hybridized references can be generated including the original two sentences:

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

4 Experiments

We will show experimental results in this section to verify the effectiveness of the extended set of hybridized reference sentences. In the experiments, multiple translations of the source language sentences are evaluated with several popular automatic evaluation metrics. The evaluation is carried out on sentence level using the original reference set and the extended reference set respectively. Finally, the Pearson's correlations between the human assessments and evaluation scores using two reference set are calculated and compared.

The multiple translations and human assessments are obtained from the dataset of the MT evaluation workshop at ACL05 (LDC2006T04) and the dataset from NistMATR08 (LDC2008E43). Table 1 & 2 describes the detail of the two datasets.

The popular automatic evaluation metrics include BLEU (Papieni et al., 2002), GTM (Melamed et al., 2003), Rouge (Lin and Och, 2004) and METEOR (Banerjee and Lavie, 2005). The syntactic trees of the reference sentences are obtained with the Stanford statistical parser (Klein 2003) for LDC2006T04 and Collins parser (Collins 1999) for LDC2008E43.

Table 3 & 4 gives out the correlations using two reference set on both datasets. The first column is the name of the used metrics. The second column is the correlations based on the original reference set. The third column is the correlations based on the extended reference set. In the experiment, the maximum length of N-gram in BLEU is 4. The exponent of GTM is 2. ROUGE uses skip-bigram with a window of nine words. And METEOR is run in "exact" mode.

Release Year	2006
Genre	Newswire
Number of segments	919
Source Language	Chinese
Target Language	English
Number of system translations	7
Number of reference translations	4
Human assessment scores	Score 1-5, adequacy & fluency

Table 1 Description of LDC2006T04

Release Year	2008
Genre	Newswire
Number of segments	249
Source Language	Arabic
Target Language	English
Number of system translations	8
Number of reference translations	4
Human assessment scores	Score 1-7, adequacy

Table 2 Description of LDC2008E43

After the hybridization, each source sentence in LDC2006T04 has 31 corresponding reference sentences in average and each source sentence in LDC2008E43 has 66 corresponding reference sentences in average. The number of the references is greatly increased. And as shown in the results, the usage of the extended reference set improves the correlations with human assessments for all the metrics in most cases except the ROUGE on LDC 2008E43.

Metric	Original	Extended
BLEU	0.3488	0.3564
GTM	0.3671	0.3681
ROUGE	0.4252	0.4325
METEOR	0.4686	0.4723

Table 3 Pearson's correlations with human assessments on sentence level on LDC2006T04

Metric	Original	Extended
BLEU	0.6092	0.6109
GTM	0.5434	0.5438
ROUGE	**0.6628**	**0.6582**
METEOR	0.7053	0.7089

Table 4 Pearson's correlations with human assessments on sentence level on LDC2008E43

The following is a real instance in the experiments from LDC2008E43:

Four original references:

- Ten churches burned down in 10 days in the American state of Alabama
- Burning of ten churches in ten days in the American state of Alabama
- Ten churches set on fire in ten days in American state of Alabama
- Torching of ten churches within ten days in American state of Alabama

Six additional references:

- Torching of ten churches in ten days in the American state of Alabama
- Torching of ten churches within ten days in the American state of Alabama
- Torching of ten churches in ten days in American state of Alabama
- Burning of ten churches within ten days in American state of Alabama
- Burning of ten churches within ten days in the American state of Alabama
- Burning of ten churches in ten days in American state of Alabama

The syntactic structure of the original references:

- (TOP (S (NPB (CD Ten) (NNS Churches)) (VP (VBN Burned) (PP (IN Down) (PP (IN in) (NP (NPB (CD 10) (NNS Days)) (PP (IN in) (NP (NPB (DT the) (NNP American) (NNP State)) (PP (IN of) (NPB (NNP Alabama)))))))))))
- (TOP (NP (NPB (NN Burning)) (PP (IN of) (NP (NPB (CD Ten) (NNS Churches)) (PP (IN in) (NP (NPB (CD Ten) (NNS Days)) (PP (IN in) (NP (NPB (DT the) (NNP American) (NNP State)) (PP (IN of) (NPB (NNP Alabama)))))))))))
- (TOP (S (NPB (CD Ten) (NNS Churches)) (VP (VB Set) (PP (IN on) (NPB (NN Fire))) (PP (IN in) (NP (NPB (CD Ten) (NNS Days)) (PP (IN in) (NP (NPB (NNP American) (NNP State)) (PP (IN of) (NPB (NNP Alabama)))))))))))
- (TOP (NP (NPB (NNP Torching)) (PP (IN of) (NP (NPB (CD Ten) (NNS Churches)) (PP (IN within) (NP (NPB (CD Ten) (NNS Days)) (PP (IN in) (NP (NPB (NNP American) (NNP State)) (PP (IN of) (NPB (NNP Alabama)))))))))))

To investigate the distribution of the equivalents we also perform several statistics about the count and the length of the syntactic nodes. In table 5, we list the information about the count of the nodes. The first row is the average words count per reference sentence. The second and third row is the count of all tree nodes and equivalent nodes in all references respectively. The fourth and fifth row is the average count of tree nodes and equivalent nodes per reference sentence respectively.

	2006T04	2008E43
Average length of reference	31.52	34.43
Total tree nodes	211231	62569
Total equivalent nodes	21807	10073
Average tree nodes	57.46	62.82
Average equivalent nodes	5.93	10.11

Table 5 Counts of the tree nodes and equivalent nodes in references.

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 pro-portion of equivalent nodes and tree nodes for each length. Figure 2 and 3 illustrate 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 4 and 5 illustrate the distribution of the proportions on two datasets respectively. The X-axis is the length of the nodes and the Y-axis is the proportion.

The investigation reveals four main messages. First, the absolute counts of the short equivalents are much more than those of long equivalents as expected. Second, the proportion of the long equivalents is greater than those of short equivalents, this clarify that the reason of large amount of short equivalents is the large amount of short tree nodes. Third, also from the proportion of view we can see that the new method comparably bias to the long equivalents. This happens because the method adopts a top-down survey of the tree. Forth, the multiple references in Arabic-English data seem to match each other better than the references

in Chinese-English data. Arabic-English references have much more equivalents than Chinese-English data and bias to long equivalents more significant.

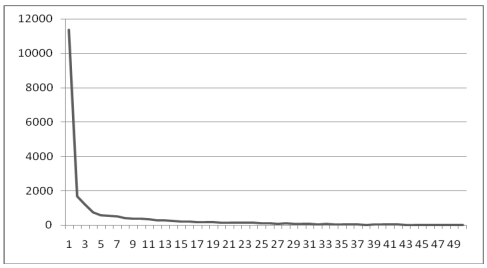


Figure 2 Distribution of absolute length of equivalent node on LDC2006T04

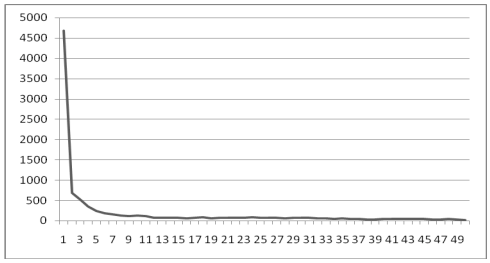


Figure 3 Distribution of absolute length of equivalent node on LDC2008E43

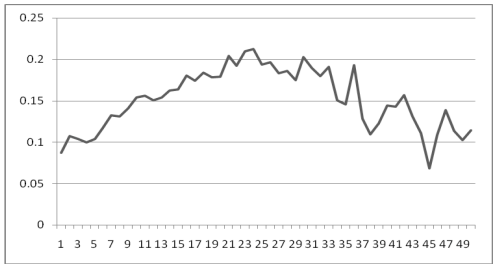


Figure 4 Distribution of length proportion of equivalent nodes on LDC2006T04

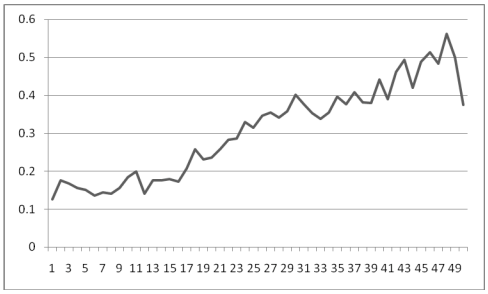


Figure 5 Distribution of length proportion of equivalent nodes on LDC2008E43

5 Discussion

The experimental results verify the positive effect of the hybridized reference for the automatic evaluation in most cases. Though the improvement of the correlations is not very significant it is stable across the metrics in various styles.

Compared with the previous works based on paraphrase and synonym the new method has three important advantages. The first is that the hybridized reference can switch the long span sub-segments beyond the words and phrases.

The second is that the switch can be per-formed in multiple levels, i.e. a sub-segment can not only be replaced as a single unit but also can be varied by replacing some child sub-segments of it. It's noticeable that the multiple level switches also make it possible to present some structural variations by means of the lexical variations. In hybridization, we can realize some structural variation between syntactic nodes by switch their parent node instead of reordering them directly.

The third advantage is that the new method needs no external knowledge which greatly facilitates the application. But this advantage also results in the main deficiency of this approach: the hybridization references cannot adopt any novel equivalents which are absent in existing references. This deficiency can be overcome by introducing the paraphrase and synonym into the syntactic hybridization.

It should be indicated that though the hybridization process generate many new references not all of the new references are reasonable.

In table 6 we compare the effect of hybridized references and manual references with more details on LDC2006T04. In the table, the first column is the contents of the references for each source sentence. "Manual" means the manual references and the number in front of it indicates how many manual references are provided. "Hybr" means the hybridized references generated from the manual references in front of the "+". The second column is the Pearson's correlations between human assessments and the BLEU scores using the corresponding reference set. Besides the set containing 4 references the other correlations are the average of the correlations based on all possible subset containing certain number of references. For example correlation of "2 Manual" is the average of the correlations based on 6 possible subset containing 2 references.

Reference Set	Correlation
1 Manual	0.2565

2 Manual	0.3057
2 Manual+ Hybr	0.3082
3 Manual	0.3316
3 Manual + Hybr	0.3369
4 Manual	0.3488
4 Manual+ Hybr	0.3564

Table 6 Pearson's correlations based on incremental reference set

As shown in the Table 6 hybridized references can improve the correlations with human assessments on different sizes of manual references set. But it also indicated that though hybridization can generate a mass of novel references the new references is always not more effective than even one additional manual references. This tells us that the quality of the hybridized references still need to be further refined.

Another message revealed by the table is that with the increase of the number of manual references the improvement of correlation made by additional manual references is decreasing. However, the improvement made by the hybridized is increasing. This happens because the number of hybridized references increases much faster than the number of manual references.

There are still several noticeable deficiencies of this work. First, it only works when there are more than two existing references. This make it cannot be used to extend the single reference in mass bilingual corpus. Second, which is also the most important one is that this method strongly focuses on the precision at the cost of recall. Though we have recognized many equivalents for each sentence but there are still many equivalents that share different context cannot be recognized. This will be our main future work. The last deficiency is the bias to the long equivalents. This problem is caused by the same reason with the second deficiency: this method define the equivalent with the same syntactic context. If two sub-nodes do not share the same parent it often have different brothers.

6 Conclusions and Future Work

In this work we present a novel method to extend the coverage of the reference set for the automatic evaluation of machine translation. The new method decomposes the existing references into sub-segments according to the syntactic structure. And then generate new reference sentences by hybridizing the equivalents of the segments which play the same syntactic role in corresponding references. In this way the new method can not only capture the equivalents of words and phrases like the other methods but also capture the equivalents of long sub-segments which are out of the capability of the other methods. Another important advantage of the new method is the no use of the external knowledge which greatly facilitates the application.

Experimental results show that with the extended reference set the state-of-the-arts automatic evaluation metrics achieve better correlation with the human assessments.

In the future work, we will relax the restriction of the equivalent definition and try to recognize more equivalents. We will also introduce the paraphrase and synonyms into our method to see further improvement. Another interesting challenge is to hybridize the equivalents in the different order and present the structural variations directly.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 60773066 and 60736014, the National High Technology Development 863 Program of China under Grant No. 2006AA010108.

References

Statanjeev Banerjee, Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Im-proved Correlation with Human Judgements. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Trans-lation and/or Summarization.

M. Collins. 1999. Head-Driven Statistical Models for Natural Language Parsing. PhD Dissertation, University of Pennsylvania.

I. Dan Melamed, Ryan Green, Joseph P. Turian, 2003, Precision and recall of machine translation, In Proceedings of HLT/NAACL 2003.

David Kauchak, Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation, In Proceedings of the NAACL 2006.

Dan Klein, Christopher Manning. 2003. Accurate Unlexicalized Parsing. In Proceedings of the 41th Meeting of the ACL, pp. 423-430.

Yves Lepage, Etienne Denoual. 2005. Automatic generation of paraphrases to be used as translation refer-

ences in objective evaluation measures of ma-chine translation, In Proceedings of the IWP 2005.

Karolina Owczarzak, Declan Groves, Josef Van Genabith ,Andy Way. 2006. Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation, In Proceedings of the Workshop on Statistical Ma-chine Translation.

Karolina Owczarzak, Josef Van Genabith, Andy Way. 2007. Dependency-Based Automatic Evaluation for Machine Translation, In Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation.

Kishore Papieni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation, In Proceedings of the 40th Meeting of the ACL.

Grazia Russo-Lassner, Jimmy Lin, Philip Resnik. 2005. Re-evaluating Machine Translation Results with Paraphrase Support, Technical Report LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57, University of Maryland, College Park, MD.

Chin-Yew Lin, Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram sta-tistics. In Proceedings of the 42th Meeting of the ACL.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating Machine Translation Results with Paraphrase Support, In Proceedings of the EMNLP 2006.