DCU Participation in WMT2013 Metrics Task

Xiaofeng Wu[†], **Hui Yu**[⋆],**Qun Liu**[†]
[†]CNGL, Dublin City University, Ireland
[⋆]ICT, Chinese Academy of Sciences, China
[†]{xfwu, qliu}@computing.dcu.ie
[⋆]yuhui@ict.ac.cn

Abstract

In this paper, we propose a novel syntactic based MT evaluation metric which only employs the dependency information in the source side. Experimental results show that our method achieves higher correlation with human judgments than BLEU, TER, HWCM and METEOR at both sentence and system level for all of the four language pairs in WMT 2010.

1 Introduction

Automatic evaluation plays a more important role in the evolution of machine translation. At the earliest stage, the automatic evaluation metrics only use the lexical information, in which, BLEU (Papineni et al., 2002) is the most popular one. BLEU is simple and effective. Most of the researchers regard BLEU as their primary evaluation metric to develop and compare MT systems. However, BLEU only employs the lexical information and cannot adequately reflect the structural level similarity. Translation Error Rate (TER) (Snover et al., 2006) measures the number of edits required to change the hypothesis into one of the references. METEOR (Lavie and Agarwal, 2007), which defines loose unigram matching between the hypothesis and the references with the help of stemming and Wordnet-looking-up, is also a lexical based method and achieves the first-class human-evaluation-correlation score. AMBER (Chen and Kuhn, 2011; Chen et al., 2012) incorporates recall, extra penalties and some text processing variants on the basis of BLEU. The main weakness of all the above lexical based methods is that they cannot adequately reflect the structural level similarity.

To overcome the weakness of the lexical based methods, many syntactic based metrics were proposed. Liu and Gildea (2005) proposed STM, a constituent tree based approach, and HWCM, a dependency tree based approach.

Both of the two methods compute the similarity between the sub-trees of the hypothesis and the reference. Owczarzak et al (2007a; 2007b; 2007c) presented a method using the Lexical-Functional Grammar (LFG) dependency tree. MAXSIM (Chan and Ng, 2008) and the method proposed by Zhu et al (2010) also employed the syntactic information in association with lexical information.With the syntactic information which can reflect structural information, the correlation with the human judgments can be improved to a certain extent.

As we know that the hypothesis is potentially noisy, and these errors expand through the parsing process. Thus the power of syntactic information could be considerably weakened.

In this paper, we attempt to overcome the shortcoming of the syntactic based methods and propose a novel dependency based MT evaluation metric. The proposed metric only employs the reference dependency tree which contains both the lexical and syntactic information, leaving the hypothesis side unparsed to avoid the error propagation. In our metric, F-score is calculated using the string of hypothesis and the dependency based n-grams which are extracted from the reference dependency tree.

Experimental results show that our method achieves higher correlation with human judgments than BLEU, HWCM, TER and METEOR at both sentence level and system level for all of the four language pairs in WMT 2010.

2 Background: HWCM

HWCM is a dependency based metric which extracts the headword chains, a sequence of words which corresponds to a path in the dependency tree, from both the hypothesis and the reference dependency tree. The score of HWCM is obtained

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 435–439,
Sofia, Bulgaria, August 8-9, 2013

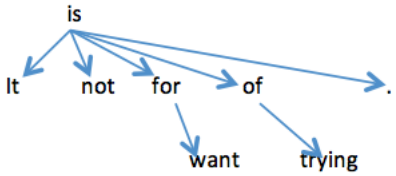

Figure 1: The dependency tree of the reference

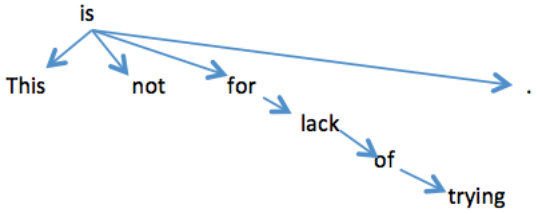

Figure 2: The dependency tree of the hypothesis

by formula (1).

$$HWCM = \frac{1}{D}\sum_{n=1}^{D}\frac{\sum_{g\in chain_n(hyp)} count_{clip}(g)}{\sum_{g\in chain_n(hyp)} count(g)} \quad (1)$$

In formula (1), D is the maximum length of the headword chain. $chain_n(hyp)$ denotes the set of the headword chains with length of n in the tree of hypothesis. $count(g)$ denotes the number of times g appears in the headword chain of the hypothesis dependency tree and $count_{clip}(g)$ denotes the clipped number of times when g appears in the the headword chain of the reference dependency trees. Clipped means that the count computed from the headword chain of the hypothesis tree should not exceed the maximum number of times when g occurs in headword chain of any single reference tree. The following are two sentences representing as reference and hypothesis, and Figure 1 and Figure 2 are the dependency trees respectively.

reference: It is not for want of trying .

hypothesis: This is not for lack of trying .

In the example above, there are 8 1-word, 7 2-word and 3 3-word headword chains in the hypothesis dependency tree. The number of 1-word and 2-word headword chains in the hypothesis tree which can match their counterparts in the reference tree is 5 and 4 respectively. The 3-word headword chains in the hypothesis dependency tree are *is for lack, for lack of* and *lack of trying*. Due to the difference in the dependency structures, they have no matches in the reference side.

3 A Novel Dependency Based MT Evaluation Method

In this new method, we calculate F-score using the string of hypothesis and the dep-n-grams which are extracted from the reference dependency tree. The new method is named DEPREF since it is a DEPendency based method only using dependency tree of REference to calculate the F-score. In DEPREF, after the parsing of the reference sentences, there are three steps below being carried out. 1) Extracting the dependency based n-gram (dep-n-gram) in the dependency tree of the reference. 2) Matching the dep-n-gram with the string of hypothesis. 3) Obtaining the final score of a hypothesis. The detail description of our method will be found in paper (Liu et al., 2013) . We only give the experiment results in this paper.

4 Experiments

Both the sentence level evaluation and the system level evaluation are conducted to assess the performance of our automatic metric. At the sentence level evaluation, Kendall's rank correlation coefficient τ is used. At the system level evaluation, the Spearman's rank correlation coefficient ρ is used.

4.1 Data

There are four language pairs in our experiments including German-to-English, Czech-to-English, French-to-English and Spanish-to-English, which are all derived from WMT2010. Each of the four language pairs consists of 2034 sentences and the references of the four language pairs are the same. There are 24 translation systems for French-to-English, 25 for German-to-English, 12 for Czech-to-English and 15 for Spanish-to-English. We parsed the reference into constituent tree by Berkeley parser and then converted the constituent tree into dependency tree by Penn2Malt [1]. Presumably, we believe that the performance will be even better if the dependency trees are manually revised.

In the experiments, we compare the performance of our metric with the widely used lexical based metrics BLEU, TER, METEOR and a dependency based metric HWCM. In order to make a fair comparison with METEOR which is known to perform best when external resources like stem and synonym are provided, we also provide results of DEPREF with external resources.

[1] http://w3.msi.vxu.se/ nivre/research/Penn2Malt.html

Metrics		Czech-English	German-English	Spanish-English	French-English
BLEU		0.2554	0.2748	0.2805	0.2197
TER		0.2526	0.2907	0.2638	0.2105
HWCM	N=1	0.2067	0.2227	0.2188	0.2022
	N=2	0.2587	0.2601	0.2408	0.2399
	N=3	0.2526	0.2638	0.2570	0.2498
	N=4	0.2453	0.2672	0.2590	0.2436
DEPREF		**0.3337**	**0.3498**	**0.3190**	**0.2656**

Table 1.A Sentence level correlations of the metrics without external resources.

Metrics	Czech-English	German-English	Spanish-English	French-English
METEOR	0.3186	0.3482	0.3258	0.2745
DEPREF	**0.3281**	**0.3606**	**0.3326**	**0.2834**

Table 1.B Sentence level correlations of the metrics with stemming and synonym.

Table 1: The sentence level correlations with the human judgments for Czech-to-English, German-to-English, Spanish-to-English and French-to-English. The number in bold is the maximum value in each column. N stands for the max length of the headword chains in HWCM in Table 1.A.

4.2 Sentence-level Evaluation

Kendall's rank correlation coefficient τ is employed to evaluate the correlation of all the MT evaluation metrics and human judgements at the sentence level. A higher value of τ means a better ranking similarity with the human judges. The correlation scores of the four language pairs and the average scores are shown in Table 1.A (without external resources) and Table 1.B (with stemming and synonym), Our method performs best when maximum length of dep-n-gram is set to 3, so we present only the results when the maximum length of dep-n-gram equals 3.

From Table 1.A, we can see that all our methods are far more better than BLEU, TER and HWCM when there is no external resources applied on all of the four language pairs. In Table 1.B, external resources is considered. DEPREF is also better than METEOR on the four language pairs. From the comparison between Table 1.A and Table 1.B, we can conclude that external resources is helpful for DEPREF on most of the language pairs. When comparing DEPREF without external resources with METEOR, we find that DEPREF obtains better results on Czech-English and German-English.

4.3 System-level Evaluation

We also evaluated the metrics with the human rankings at the system level to further investigate the effectiveness of our metrics. The matching of the words in DEPREF is correlated with the position of the words, so the traditional way of computing system level score, like what BLEU does, is not feasible for DEPREF. Therefore, we resort to the way of adding the sentence level scores together to obtain the system level score. At system level evaluation, we employ Spearman's rank correlation coefficient ρ. The correlations of the four language pairs and the average scores are shown in Table 2.A (without external resources) and Table 2.B (with stem and synonym).

From Table 2.A, we can see that the correlation of DEPREF is better than BLEU, TER and HWCM on German-English, Spanish-English and French-English. On Czech-English, our metric DEPREF is better than BLEU and TER. In Table 2.B (with stem and synonym), DEPREF obtains better results than METEOR on all of the language pairs except one case that DEPREF gets the same result as METEOR on Czech-English. When comparing DEPREF without external resources with METEOR, we can find that DEPREF gets better result than METEOR on Spanish-English and French-English.

From Table 1 and Table 2, we can conclude that, DEPREF without external resources can obtain comparable result with METEOR, and DEPREF with external resources can obtain better results than METEOR. The only exception is that at the system level evaluation, Czech-English's best score is abtained by HWCM. Notice that there are only 12 systems in Czech-English, which means there are only 12 numbers to be sorted, we believe

Metrics		Czech-English	German-English	Spanish-English	French-English
BLEU		0.8400	0.8808	0.8681	0.8391
TER		0.7832	0.8923	0.9033	0.8330
HWCM	N=1	0.8392	0.7715	0.7231	0.6730
	N=2	0.8671	0.8600	0.7670	0.8026
	N=3	**0.8811**	0.8831	0.8286	0.8209
	N=4	**0.8811**	0.9046	0.8242	0.8148
DEPREF		0.8392	**0.9238**	**0.9604**	**0.8687**

Table 2.A System level correlations of the metrics without external resources.

Metrics	Czech-English	German-English	Spanish-English	French-English
METEOR	**0.8392**	0.9269	0.9516	0.8652
DEPREF	**0.8392**	**0.9331**	**0.9692**	**0.8730**

Table 2.B System level correlations of the metrics with stemming and synonym.

Table 2: The system level correlations with the human judgments for Czech-to-English, German-to-English, Spanish-to-English and French-to-English. The number in bold is the maximum value in each column. N stands for the max length of the headword chains in HWCM in Table 2.A.

the spareness issure is more serious in this case.

5 Conclusion

In this paper, we propose a new automatic MT evaluation method DEPREF. The experiments are carried out at both sentence-level and system-level using four language pairs from WMT 2010. The experiment results indicate that DEPREF achieves better correlation than BLEU, HWCM, TER and METEOR at both sentence level and system level.

References

Yee Seng Chan and Hwee Tou Ng. 2008. Maxsim: A maximum similarity metric for machine translation evaluation. In *Proceedings of ACL-08: HLT*, pages 55–62.

Boxing Chen and Roland Kuhn. 2011. Amber: A modified bleu, enhanced ranking metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 71–77, Edinburgh, Scotland, July. Association for Computational Linguistics.

Boxing Chen, Roland Kuhn, and George Foster. 2012. Improving amber, an mt evaluation metric. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 59–63, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32.

Q. Liu, H. Yu, X. Wu, J. Xie, Y. Lu, and S. Lin. 2013. A Novel Dependency Based MT Evaluation Method. *Under Review*.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007a. Dependency-based automatic evaluation for machine translation. In *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, SSST '07, pages 80–87, Stroudsburg, PA, USA. Association for Computational Linguistics.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007b. Evaluating machine translation with lfg dependencies. *Machine Translation*, 21(2):95–119, June.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007c. Labelled dependencies in machine translation evaluation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 104–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Junguo Zhu, Muyun Yang, Bo Wang, Sheng Li, and Tiejun Zhao. 2010. All in strings: a powerful string-based automatic mt evaluation metric with multiple granularities. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1533–1540, Stroudsburg, PA, USA. Association for Computational Linguistics.