CASICT-DCU Participation in WMT2015 Metrics Task

Hui Yu[†] **Qingsong Ma**[†] **Xiaofeng Wu**[‡] **Qun Liu**[‡†]
[†]Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
[‡]ADAPT Centre, School of Computing, Dublin City University
{yuhui,maqingsong}@ict.ac.cn
{xiaofengwu}@computing.dcu.ie
{liuqun}@ict.ac.cn

Abstract

Human-designed sub-structures are required by most of the syntax-based machine translation evaluation metrics. In this paper, we propose a novel evaluation metric based on dependency parsing model, which does not need this human involvement. Experimental results show that the new single metric gets better correlation than METEOR on system level and is comparable with it on sentence level. To introduce more information, we combine the new metric with many other metrics. The combined metric obtains state-of-the-art performance on both system level evaluation and sentence level evaluation on WMT 2014.

1 Introduction

Automatic evaluation metrics play an important role in machine translation research. At present, most of the automatic evaluation metrics evaluate the translation quality by comparing the similarity between the hypothesis and the reference.

The lexicon-based metrics can only use lexical information, such as BLEU (Papineni et al., 2002), NIST(Doddington, 2002) and METEOR (Lavie and Agarwal, 2007). To evaluate the hypothesis on syntactic level, some researchers proposed the syntax-based metrics. Liu and Gildea (2005) proposed a constituent-tree-based metric STM and a dependency-tree-based metric HWCM. The syntax-based metric proposed by Owczarzak et al (2007) uses the Lexical-Functional Grammar (LFG) dependency tree. Some metrics introduce the syntactic information on the basis of lexical information, such as MAXSIM (Chan and Ng, 2008) and the metric proposed by Zhu et al. (2010). These metrics evaluate the syntactic similarity by comparing the sub-structures extracted from the trees of hypothesis and reference. To avoid parsing the hypothesis in order to prevent translation error propagation, some researchers propose a kind of syntax-based evaluation metric which only uses the tree of reference, such as BLEUÂTRE (Mehay and Brew, 2007) and RED (Yu et al., 2014).

The syntax-based metrics either use the sub-structures of both the reference and the hypothesis tree, or only use that on the reference side. Therefore, for these metrics, sub-structures designed by human are required. In this paper, we propose a novel dependency-parsing-model-based metric in the view of dependency tree generation, which completely avoids this human involvement. A dependency parsing model is trained by the reference dependency tree, through which we can obtain the dependency tree of the hypothesis and the corresponding score. The syntactic similarity between the hypothesis and the reference can be evaluated by this score. In order to obtain the lexicon similarity, we also introduce the unigram F-score to the new metric. The experimental results show that the new metric gets the state-of-the-art performance in the single metrics on system level evaluation, and gets the comparable correlation with METEOR on sentence level evaluation. We also propose a combined metric[1] which combines the new metric with many other metrics together. The combined metric obtains state-of-the-art performance on both system level and sentence level.

The remainder of this paper is organized as follows: Section 2 describes the dependency-parsing-model-based metric; Section 3 presents the combined metric; Section 4 gives the experiment results; Conclusions are discussed in Section 5.

[1] Combined metrics directly use the scores of many kinds of metrics, such as BLEU, TER, METEOR and some syntax-based metrics. For the metrics using different kinds of information types (lexicon, syntax and semantic information) as features, we still think they are single metrics, because they don't use the score of other metrics.

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 417–421,
Lisboa, Portugal, 17-18 September 2015. ©2015 Association for Computational Linguistics.

2 DPMF: Evaluation Metric Based on Dependency Parsing Model

We evaluate the syntactic similarity via the **D**ependency **P**aring **M**odel score of hypothesis and evaluate the lexical similarity via the unigram **F**-score. So we name the new metric as DPMF.

There are four steps to obtain the dependency parsing model score of the hypothesis. 1) Obtain the reference dependency tree which can be generated by the automatic parsing tools or labeled by human. 2) Train a dependency parsing model using the reference dependency tree. 3) Parse the hypothesis using the dependency parsing model and get the probability of the hypothesis dependency tree. 4) Normalize the probability of the hypothesis dependency tree. We define the normalized probability of the hypothesis dependency tree as the dependency parsing model score. After obtaining the dependency parsing model score of a hypothesis, we multiply this score by unigram F-score to get the final score of DPMF. The detailed description of our metric will be found in paper Yu et al. (2015a). We only give the experiment results in this paper.

3 DPMF$_{comb}$: A Combined Evaluation Metric

From the published results of WMT 2014, we can see that the combined metrics such as DISCOTK-PARTY (Joty et al., 2014) and UPC-STOUT (Gonzàlez et al., 2014) obtained great success which can make use of many single metrics. In most of the cases, combined metrics can obtain good correlations, so we also propose a combined metric which combines DPMF with some other single metrics. The combined metric is named as DPMF$_{comb}$ and it involves DPMF, REDp, ENTFp[2] and some metrics included in the open source toolkit Asiya[3].

We introduce REDp, ENTFp and Asiya briefly in the rest of this section.

3.1 REDp

RED (Yu et al., 2014) employs the reference dependency tree which contains both the lexical and syntactic information, leaving the hypothesis side unparsed to avoid error propagation. The score of RED is obtained using F-score. The precision and recall are calculated using the dependency tree of the reference and the string of the hypothesis. To extend the limited reference, they introduce some linguistic resources into RED and propose a new version REDp, which is employed in our combined metric. We merge the extended version REDp into our combined metric.

3.2 ENTFp

The widely-used lexicon-based evaluation metrics cannot adequately reflect the fluency of the translations. The n-gram-based metrics, like BLEU, limit the maximum length of matched fragments to N and cannot catch the matched fragments longer than N, so they can only reflect the fluency indirectly. METEOR, which is not limited by n-gram, uses the number of matched chunks but it does not consider the length of each chunk. To avoid this defect, we propose an entropy-based method ENTF, which is a metric by introducing unigram F-score on the base of ENT (Yu et al., 2015b). ENT aims at reflecting the fluency of translations through the distribution of matched words, while the unigram F-score can evaluate the accuracy. We introduce stem, synonym and paraphrase into ENTF to extend the limited number of reference and name it as ENTFp.

3.3 Asiya

We use Asiya MT evaluation toolkit (Giménez and Màrquez, 2010) to produce the score of many metrics, which can be used in DPMF$_{comb}$. Asiya provides a rich set of specialized similarity metrics that use different level of linguistic information, namely lexical, syntactic and semantic.

In our experiment, we calculate scores of the default metric set provided by Asiya. For the into-English language pairs, the default metric set contains 55 metrics, including lexicon-based metrics, syntax-based metrics and semantic-based metrics. The weights of all these 55 scores together with the scores of DPMF, REDp and ENTFp are trained with SVM-rank[4].

4 Experiments

To evaluate the performance of DPMF and DPMF$_{comb}$, we carry out the experiments on both system level evaluation and sentence level evaluation. In this section, we first describe the data sets

[2] The source code of DPMF, REDp and ENTFp can be found in http://github.com/YuHui0117/AMTE

[3] http://asiya-faust.cs.upc.edu/

[4] http://www.cs.cornell.edu/People/tj/svm_light/svm_rank.html

and the baseline metrics in the experiments, and then give and analyse the experimental results.

4.1 Data

We use the data from the WMT 2014 evaluation campaign as test data. The language pairs are Czech-to-English, German-to-English, French-to-English and Russian-English. The number of translation systems for each language pair are shown in Table 1.

data	cs-en	de-en	fr-en	ru-en
WMT2014	5	13	8	13

Table 1: The number of translation systems for each language pair on WMT 2014. cs-en means Czech-to-English. de-en means German-to-English. fr-en means French-to-English. ru-en means Russian-to-English.

DPMF$_{comb}$ is a combined metric which includes 58 single metrics. The training data used to train the weight of each single metric are the English-targeted language pairs in WMT 2012 and WMT 2013.

4.2 Baseline

The baselines are the widely-used lexicon-based metrics, such as BLEU[5], TER[6] and METEOR[7]. In addition, according to the published results of WMT 2014, we also give the correlation of the metric with the best performance on average, DISCOTK-PARTY-TUNED (Joty et al., 2014), which is a combined metric including many kinds of other metrics. For fairness, we also give the result of the metric with the best performance on average in the single metrics, VERTA-W(Comelles and Atserias, 2014) on system level and BEER (Stanojevic and Sima'an, 2014) on sentence level respectively. For our combined metric, to evaluate the effect of adding DPMF, REDp and ENTFp, we also give the correlation of the metric only combining the single metrics in Asiya.

4.3 System Level Correlation

To verify the effectiveness of DPMF, we carry out the system level experiments on WMT 2014. To evaluate the correlation with human judges, Spearman's rank correlation coefficient ρ is used. ρ is calculated using Formula (1).

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{1}$$

d_i is the difference between the human rank and metrics rank for system i. n is the number of systems.

We give the system level correlations for every metric in Table 2. From Table 2, we can see that DPMF is better than BLEU and TER on all the language pairs. The correlation of DPMF is also better than METEOR and the best single metric VERTA-W on average. But DPMF is lower than the combined metrics DISCOTK-PARTY-TUNED and Asiya. After combining DPMF with other metrics, DPMF$_{comb}$ obtains better correlations than DISCOTK-PARTY-TUNED and Asiya on average. We can see that DPMF$_{comb}$ obtains state-of-the-art performance on system level. From the comparison between DPMF$_{comb}$ and Asiya, we can see that adding DPMF, REDp and ENTFp into the combined metric is useful on system level.

4.4 Sentence Level Correlation

To further evaluate the performance of DPMF and DPMF$_{comb}$, we carry out the experiments on sentence level. On sentence level, Kendall's τ correlation coefficient is used. τ is calculated using the following equation.

$$\tau = \frac{\text{num\_con\_pairs} - \text{num\_dis\_pairs}}{\text{num\_con\_pairs} + \text{num\_dis\_pairs}}$$

num_con_pairs is the number of concordant pairs and num_dis_pairs is the number of disconcordant pairs.

Table 3 gives the correlations of all the metrics. We can see that DPMF is better than BLEU on each language pair and it is comparable with METEOR on average. The stat-of-the-art performance on sentence level is obtained after combining DPMF with other metrics, namely, DPMF$_{comb}$, which outperforms the combined metrics DISCO-PARTY-TUNED and Asiya. From the comparison between DPMF$_{comb}$ and Asiya, we can see that adding DPMF, REDp and ENTFp into the combined metric is useful on sentence level.

[5] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl

[6] http://www.cs.umd.edu/~snover/tercom

[7] http://www.cs.cmu.edu/~alavie/METEOR/download/meteor-1.4.tgz

metrics	cs-en	de-en	fr-en	ru-en	average
TER	.976	.775	.952	.809	.878
BLEU	.909	.832	.952	.789	.871
METEOR	.980	.927	.975	.805	.922
DISCOTK-PARTY-TUNED	.975	.943	.977	.870	.941
VERTA-W	.934	.867	.959	.848	.902
Asiya	.954	.936	**.978**	.871	.935
DPMF	**.999**	.920	.967	.832	.930
DPMF$_{comb}$	.974	**.950**	**.978**	**.872**	**.944**

Table 2: System level correlations on WMT 2014. Asiya represents the combined metric only using the metrics in Asiya. The value in bold is the best result in each column. *average* stands for the average result of all the language pairs for each metric on WMT 2014.

metrics	cs-en	de-en	fr-en	ru-en	average
BLEU	.216	.259	.367	.256	.275
METEOR	.282	.334	.406	.329	.338
BEER	.284	.337	.417	.333	.343
DISCOTK-PARTY-TUNED	.328	.380	.433	.355	.374
Asiya	**.333**	.388	.437	.355	.378
DPMF	.283	.332	.404	.324	.336
DPMF$_{comb}$	.332	**.398**	**.443**	**.364**	**.384**

Table 3: Sentence level correlations on WMT 2014. Asiya represents the combined metric only using the metrics in Asiya. The value in bold is the best result in each column. *average* stands for the average result of all the language pairs for each metric on WMT 2014.

5 Conclusion

In this paper, we propose a new dependency-parsing-model-based metric DPMF and a combined metric DPMF$_{comb}$. DPMF evaluates the syntactic similarity through the dependency parsing model and evaluates the lexical similarity by unigram F-score. Experimental results show that the correlation of DPMF is better than BLEU, TER, METEOR and VERTA-w on system level. On sentence level, DPMF is better than BLEU, and comparable with METEOR. After combining DPMF with other metrics, DPMF$_{comb}$ obtains the state-of-the-art performance on both system level and sentence level on WMT 2014.

Acknowledgments

This research is supported by China's NSFC grant 61379086 and the European Union Horizon 2020 Programme (H2020) under grant agreement no. 645452 (QT21). The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

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.

Elisabet Comelles and Jordi Atserias. 2014. Verta participation in the wmt14 metrics task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 368–375, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jesús Giménez and Lluís Màrquez. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Meritxell Gonzàlez, Alberto Barrón-Cedeño, and Lluís Màrquez. 2014. Ipa and stout: Leveraging linguistic and source-based features for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 394–401, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2014. Discotk: Using discourse structure for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 402–408, Baltimore, Maryland, USA, June. 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.

Dennis Mehay and Chris Brew. 2007. BLEUÂTRE: Flattening Syntactic Dependencies for MT Evaluation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

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

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.

Milos Stanojevic and Khalil Sima'an. 2014. Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Hui Yu, Xiaofeng Wu, Jun Xie, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2014. Red: A reference dependency based mt evaluation metric. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2042–2051, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Hui Yu, Xiaofeng Wu, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2015a. An Automatic Machine Translation Evaluation Metric Based on Dependency Parsing Model. *ArXiv e-prints*, August.

Hui Yu, Xiaofeng Wu, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2015b. Improve the Evaluation of Translation Fluency by Using Entropy of Matched Sub-segments. *ArXiv e-prints*, August.

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.