RED: DCU-CASICT Participation in WMT2014 Metrics Task

Xiaofeng Wu[†], **Hui Yu**[*], **Qun Liu**[†*]
[†]CNGL Centre for Global Intelligent Content
School of Computing, Dublin City University
Dublin 9, Ireland
[*]Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
{xiaofengwu,qliu}@computing.dcu.ie, yuhui@ict.ac.cn

Abstract

Based on the last year's DCU-CASIST participation on WMT metrics task, we further improve our model in the following ways: 1) parameter tuning 2) support languages other than English. We tuned our system on all the data of WMT 2010, 2012 and 2013. The tuning results as well as the WMT 2014 test results are reported.

1 Introduction

Automatic evaluation plays a more and more important role in the evolution of machine translation. There are roughly two categories can be seen: namely lexical information based and structural information based.

1) Lexical information based approaches, among which, BLEU (**?**), Translation Error Rate (TER) (**?**) and METEOR (**?**) are the most popular ones and, with simplicity as their merits, cannot adequately reflect the structural level similarity.

2) A lot of structural level based methods have been exploited to overcome the weakness of the lexical based methods, including Syntactic Tree Model(STM)(**?**), a constituent tree based approach, and Head Word Chain Model(HWCM)(**?**), a dependency tree based approach. Both of the methods compute the similarity between the sub-trees of the hypothesis and the reference. Owczarzak et al (**?**; **?**; **?**) presented a method using the Lexical-Functional Grammar (LFG) dependency tree. MAXSIM (**?**) and the method proposed by Zhu et al (**?**) also employed the syntactic information in association with lexical information. As we know that the hypothesis is potentially noisy, and these errors are enlarged through the parsing process. Thus the power of syntactic information could be considerably weakened.

In this paper, based on our attempt on WMT metrics task 2013 (**?**), we propose a metrics named RED (REference Dependency based automatic evaluation method). Our metrics employs only the reference dependency tree which contains both the lexical and syntactic information, leaving the hypothesis side unparsed to avoid error propagation.

2 Parameter Tuning

In RED, we use $F\text{-}score$ as our final score. $F\text{-}score$ is calculated by Formula (1), where α is a value between 0 and 1.

$$F\text{-}score = \frac{precision \cdot recall}{\alpha \cdot precision + (1-\alpha) \cdot recall} \quad (1)$$

The dependency tree of the reference and the string of the translation are used to calculate the precision and recall. In order to calculate precision, the number of the dep-ngrams (the ngrams obtained from dependency tree[1]) should be given, but there is no dependency tree for the translation in our method. We know that the number of dep-ngrams has an approximate linear relationship with the length of the sentence, so we use the length of the translation to replace the number of the dep-ngrams in the translation dependency tree. Recall can be calculated directly since we know the number of the dep-ngrams in the reference. The precision and recall are computed as follows.

$$precision_n = \frac{\sum_{d \in D_n} p_{(d,hyp)}}{len_h}$$

$$recall_n = \frac{\sum_{d \in D_n} p_{(d,hyp)}}{count_{n(ref)}}$$

D_n is the set of dep-ngrams with the length of n. len_h is the length of the translation. $count_{n(ref)}$ is the number of the dep-ngrams with the length of n in the reference. $p_{(d,hpy)}$ is 0 if there is no match and a positive number between 0 and 1 otherwise(**?**).

[1]We define two types of dep-ngrams: 1) the head word chain(**?**); 2) fix-floating(**?**))

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 420–425,
Baltimore, Maryland USA, June 26–27, 2014.

The final score of RED is achieved using Formula (2), in which a weighted sum of the dep-ngrams' $F\text{-}score$ is calculated. w_{ngram} ($0 \leq w_{ngram} \leq 1$) is the weight of dep-ngram with the length of *n*. $F\text{-}score_n$ is the $F\text{-}score$ for the dep-ngrams with the length of *n*.

$$RED = \sum_{n=1}^{N}(w_{ngram} \times F\text{-}score_n) \quad (2)$$

Other parameters to be tuned includes:

- Stem and Synonym

  Stem(**?**) and synonym (WordNet[2]) are introduced with the following three steps. First, we obtain the alignment with METEOR Aligner (**?**) in which not only exact match but also stem and synonym are considered. We use stem, synonym and exact match as the three match modules. Second, the alignment is used to match for a dep-ngram. We think the dep-ngram can match with the translation if the following conditions are satisfied. 1) Each of the words in the dep-ngram has a matched word in the translation according to the alignment; 2) The words in dep-ngram and the matched words in translation appear in the same order; 3) The matched words in translation must be continuous if the dep-ngram is a fixed-floating ngram. At last, the match module score of a dep-ngram is calculated according to Formula (3). Different match modules have different effects, so we give them different weights.

  $$s_{mod} = \frac{\sum_{i=1}^{n} w_{m_i}}{n}, \ 0 \leq w_{m_i} \leq 1 \quad (3)$$

  m_i is the match module (exact, stem or synonym) of the ith word in a dep-ngram. w_{m_i} is the match module weight of the ith word in a dep-ngram. n is the number of words in a dep-ngram.

- Paraphrase

  When introducing paraphrase, we don't consider the dependency tree of the reference, because paraphrases may not be contained in the head word chain and fixed-floating structures. Therefore we first obtain the alignment with METEOR Aligner, only considering paraphrase; Then, the matched paraphrases are extracted from the alignment and defined as paraphrase-ngram. The score of a paraphrase is $1 \times w_{par}$, where w_{par} is the weight of paraphrase-ngram.

- Function word

  We introduce a parameter w_{fun} ($0 \leq w_{fun} \leq 1$) to distinguish function words and content words. w_{fun} is the weight of function words. The function word score of a dep-ngram or paraphrase-ngram is computed according to Formula (4).

  $$s_{fun} = \frac{C_{fun} \times w_{fun} + C_{con} \times (1 - w_{fun})}{C_{fun} + C_{con}} \quad (4)$$

  C_{fun} is the number of function words in the dep-ngram or paraphrase-ngram. C_{con} is the number of content words in the dep-ngram or paraphrase-ngram.

$$REDp = \sum_{n=1}^{N}(w_{ngram} \times F\text{-}score_{pn}) \quad (5)$$

$$F\text{-}score_p = \frac{precision_p \cdot recall_p}{\alpha \cdot precision_p + (1 - \alpha) \cdot recall_p} \quad (6)$$

$precision_p$ and $recall_P$ in Formula (6) are calculated as follows.

$$precision_p = \frac{score_{par_n} + score_{dep_n}}{len_h}$$

$$recall_p = \frac{score_{par_n} + score_{dep_n}}{count_n(ref) + count_n(par)}$$

len_h is the length of the translation. $count_{n(ref)}$ is the number of the dep-ngrams with the length of n in the reference. $count_n(par)$ is the number of paraphrases with length of n in reference. $score_{par_n}$ is the match score of paraphrase-ngrams with the length of *n*. $score_{dep_n}$ is the match score of dep-ngrams with the length of *n*. $score_{par_n}$ and $score_{dep_n}$ are calculated as follows.

$$score_{par_n} = \sum_{par \in P_n} (1 \times w_{par} \times s_{fum})$$

$$score_{dep_n} = \sum_{d \in D_n} (p_{(d,hyp)} \times s_{mod} \times s_{fun})$$

[2]http://wordnet.princeton.edu/

Metrics		BLEU	TER	HWCM	METEOR	RED	RED-sent	RED-syssent
WMT 2010	cs-en	0.255	0.253	0.245	0.319	0.328	**0.342**	**0.342**
	de-en	0.275	0.291	0.267	0.348	0.361	0.371	**0.375**
	es-en	0.280	0.263	0.259	0.326	0.333	0.344	**0.347**
	fr-en	0.220	0.211	0.244	0.275	0.283	**0.329**	0.328
	ave	0.257	0.254	0.253	0.317	0.326	0.346	**0.348**
WMT 2012	cs-en	0.157	-	0.158	0.212	0.165	0.218	0.212
	de-en	0.191	-	0.207	0.275	0.218	**0.283**	0.279
	es-en	0.189	-	0.203	0.249	0.203	0.255	**0.256**
	fr-en	0.210	-	0.204	0.251	0.221	0.250	**0.253**
	ave	0.186	-	0.193	0.246	0.201	**0.251**	0.250
WMT 2013	cs-en	0.199	-	0.153	**0.265**	0.228	0.260	0.256
	de-en	0.220	-	0.182	0.293	0.267	**0.298**	0.297
	es-en	0.259	-	0.220	0.324	0.312	**0.330**	0.326
	fr-en	0.224	-	0.194	0.264	0.257	**0.267**	0.266
	ru-en	0.162	-	0.136	0.239	0.200	**0.262**	0.225
	ave	0.212	-	0.177	0.277	0.252	**0.283**	0.274
WMT 2014	hi-en	-	-	-	**0.420**	-	0.383	0.386
	de-en	-	-	-	0.334	-	0.336	**0.338**
	cs-en	-	-	-	0.282	-	**0.283**	**0.283**
	fr-en	-	-	-	**0.406**	-	0.403	0.404
	ru-en	-	-	-	**0.337**	-	0.328	0.329
	ave	-	-	-	**0.355**	-	0.347	0.348

Table 1: Sentence level correlations tuned on WMT 2010, 2012 and 2013; tested on WMT 2014. The value in bold is the best result in each raw. *ave* stands for the average result of the language pairs on each year. RED stands for our untuned system, RED-sent is G.sent.2, RED-syssent is G.sent.1

P_n is the set of paraphrase-ngrams with the length of n. D_n is the set of dep-ngrams with the length of n.

There are totally nine parameters in RED. We tried two parameter tuning strategies: Genetic search algorithm (**?**) and a Grid search over two subsets of parameters. The results of Grid search is more stable, therefore our final submission is based upon Grid search. We separate the 9 parameters into two subsets. When searching Subset 1, the parameters in Subset 2 are fixed, and vice versa. Several iterations are executed to finish the parameter tuning process. For system level coefficient score, we set two optimization goals: G.sys.1) to maximize the sum of Spearman's ρ rank correlation coefficient on system level and Kendall's τ correlation coefficient on sentence level or G.sys.2) only the former; For sentence level coefficient score, we also set two optimization goals: G.sent.1) the same as G.sys.1, G.sent.2) only the latter part of G.sys.1.

3 Experiments

In this section we report the experimental results of RED on the tuning set, which is the combination of WMT2010, WMT2012 and WMT2013 data set, as well as the test results on the WMT2014. Both the sentence level evaluation and the system level evaluation are conducted to assess the performance of our automatic metrics. 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.

3.1 Data

There are four language pairs in WMT2010 and WMT2012 including German-English, Czech-English, French-English and Spanish-English. For WMT2013, except these 4 language pairs, there is also Russian-English. As the test set, WMT 2014 has also five language pairs, but the organizer removed Spanish-English and replace it with Hindi-English. For into-English tasks, we parsed the En-

Metrics		BLEU	TER	HWCM	METEOR	RED	RED-sys	RED-syssent
WMT 2010	cs-en	0.840	0.783	0.881	0.839	0.839	**0.937**	0.881
	de-en	0.881	0.892	0.905	0.927	0.933	**0.95**	0.948
	es-en	0.868	0.903	0.824	0.952	**0.969**	0.965	**0.969**
	fr-en	0.839	0.833	0.815	0.865	0.873	0.875	**0.905**
	ave	0.857	0.852	0.856	0.895	0.903	**0.931**	0.925
WMT 2012	cs-en	0.886	0.886	0.943	0.657	**1**	**1**	**1**
	de-en	0.671	0.624	0.762	0.885	0.759	0.935	**0.956**
	es-en	0.874	0.916	0.937	0.951	0.951	**0.965**	0.958
	fr-en	0.811	0.821	0.818	0.843	0.818	**0.871**	0.853
	ave	0.810	0.811	0.865	0.834	0.882	**0.942**	0.941
WMT 2013	cs-en	0.936	0.800	0.818	0.964	0.964	**0.982**	0.972
	de-en	0.895	0.833	0.816	0.961	0.951	0.958	**0.978**
	es-en	0.888	0.825	0.755	0.979	0.930	**0.979**	0.965
	fr-en	0.989	0.951	0.940	0.984	0.989	**0.995**	0.984
	ru-en	0.670	0.581	0.360	0.789	0.725	**0.847**	0.821
	ave	0.875	0.798	0.737	0.834	0.935	**0.952**	0.944
WMT 2014	hi-en	**0.956**	0.618	-	0.457	-	0.676	0.644
	de-en	0.831	0.774	-	0.926	-	0.897	**0.909**
	cs-en	0.908	0.977	-	0.980	-	0.989	**0.993**
	fr-en	0.952	0.952	-	0.975	-	**0.981**	0.980
	ru-en	0.774	0.796	-	0.792	-	**0.803**	0.797
	ave	**0.826**	0.740	-	0.784	-	0.784	0.770

Table 2: System level correlations tuned on WMT 2010, 2012 and 2013, tested on 2014. The value in bold is the best result in each raw. *ave* stands for the average result of the language pairs on each year. RED stands for our untuned system, RED-sys is G.sys.2, RED-syssent is G.sys.1

Metrics		BLEU	TER	METEOR	RED	RED-sent	RED-syssent
WMT 2010	en-fr	0.33	0.31	0.369	0.338	**0.390**	0.369
	en-de	0.15	0.08	0.166	0.141	**0.214**	0.185
WMT 2012	en-fr	-	-	0.26	0.171	**0.273**	0.266
	en-de	-	-	0.180	0.129	**0.200**	0.196
WMT 2013	en-fr	-	-	0.236	0.220	0.237	**0.239**
	en-de	-	-	0.203	0.185	0.215	**0.219**
WMT 2014	en-fr	-	-	0.278	-	**0.297**	0.293
	en-de	-	-	0.233	-	**0.236**	0.229

Table 3: Sentence level correlations tuned on WMT 2010, 2012 and 2013, and tested on 2014. The value in bold is the best result in each raw. RED stands for our untuned system, RED-sent is G.sent.2, RED-syssent is G.sent.1

glish reference into constituent tree by Berkeley parser and then converted the constituent tree into dependency tree by Penn2Malt [3]. We also conducted English-to-French and English-to-German experiments. The German and French dependency parser we used is Mate-Tool [4].

In the experiments, we compare the performance of our metric with the widely used lexical based metrics BLEU, TER, METEOR and a dependency based metrics HWCM. The results of RED are provided with exactly the same external resources like METEOR. The results of BLEU, TER and METOER are obtained from official report of WMT 2010, 2012, 2013 and 2014 (if they

[3]http://w3.msi.vxu.se/ nivre/research/Penn2Malt.html
[4]https://code.google.com/p/mate-tools/

Metrics		BLEU	TER	METEOR	RED	RED-sys	RED-syssent
WMT 2010	en-fr	0.89	0.89	0.912	0.881	**0.932**	0.928
	en-de	0.66	0.65	0.688	0.657	**0.734**	**0.734**
WMT 2012	en-fr	0.80	0.69	0.82	0.639	**0.914**	**0.914**
	en-de	0.22	**0.41**	0.180	0.143	0.243	0.243
WMT 2013	en-fr	0.897	0.912	0.924	0.914	0.931	**0.936**
	en-de	0.786	0.854	**0.879**	0.85	0.8	0.8
WMT 2014	en-fr	0.934	**0.953**	0.940	-	0.942	0.943
	en-de	0.065	**0.163**	0.128	-	0.047	0.047

Table 4: System level correlations for English to Franch and German, tuned on WMT 2010, 2012 and 2013; tested on WMT 2014. The value in bold is the best result in each raw. RED stands for our untuned system, RED-sys is G.sys.2, RED-syssent is G.sys.1

are available). The experiments of HWCM is performed by us.

3.2 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 are shown in Table 1. 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, we can see that: firstly, parameter tuning improve performance significantly on all the three tuning sets; secondly, although the best scores in the column RED-sent are much more than RED-syssent, but the outperform is very small, so by setting these two optimization goals, RED can achieve comparable performance; thirdly, without parameter tuning, RED does not perform well on WMT 2012 and 2013, and even with parameter tuning, RED does not outperform METEOR as much as WMT 2010; lastly, on the test set, RED does not outperform METEOR.

3.3 System-level Evaluation

We also evaluated the RED scores with the human rankings at the system level to further investigate the effectiveness of our metrics. The matching of the words in RED 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 RED. 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 and the average scores are shown in Table 2.

From Table 2, we can see similar trends as in Table 1 with the following difference: firstly, without parameter tuning, RED perform comparably with METEOR on all the three tuning sets; secondly, on test set, RED also perform comparably with METEOR. thirdly, RED perform very bad on Hindi-English, which is a newly introduced task this year.

3.4 Evaluation of English to Other Languages

We evaluate both sentence level and system level score of RED on English to French and German. The reason we only conduct these two languages are that the dependency parsers are more reliable in these two languages. The results are shown in Table 3 and 4.

From Table 3 and 4 we can see that the tuned version of RED still perform slightly better than METEOR with the only exception of system level en-de.

4 Conclusion

In this paper, based on the last year's DCU-CASICT submission, we further improved our method, namely RED. The experiments are carried out at both sentence-level and system-level using to-English and from-English corpus. The experiment results indicate that although RED achieves better correlation than BLEU, HWCM, TER and comparably performance with METEOR at both sentence level and system level, the performance is not stable on all language pairs, such as the sentence level correlation score of Hindi to

English and the system level score of English to German. To further study the sudden diving of the performance is our future work.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of the CNGL Centre for Global Intelligent Content (www.cngl.ie) at Dublin City University and National Natural Science Foundation of China (Grant 61379086).

References

Ergun Biçici and Deniz Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 272–283. Association for Computational Linguistics.

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.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 85–91, 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.

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.

Martin F Porter. 2001. Snowball: A language for stemming algorithms.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2010. String-to-dependency statistical machine translation. *Computational Linguistics*, 36(4):649–671.

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.

Xiaofeng Wu, Hui Yu, and Qun Liu. 2013. Dcu participation in wmt2013 metrics task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

H. Yu, X. Wu, Q. Liu, and S. Lin. 2014. RED: A Reference Dependency Based MT Evaluation Metric. In *To be published*.

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.