Insight into Multiple References in an MT Evaluation Metric

Ying Qin, Lucia Specia

Beijing Foreign Studies University, 100089, China,
qinying@bfsu.edu.cn
University of Sheffield, S10 2TN, UK
l.specia@sheffield.ac.uk

Abstract. Current evaluation metrics in machine translation (MT) make poor use of multiple reference translations. In this paper we focus on the METEOR metric to gain in-depth insights into how best multiple references can be exploited. Results on five score selection strategies reveal that it is not always wise to choose the best (closest to MT) reference to generate the candidate score. We also propose two weighting approaches by taking into account the recurring information among references. The modified METEOR scores significantly increase the correlation with human judgments on accuracy and fluency evaluation at system level.

Keywords: machine translation evaluation, METEOR metric, multiple references

1 Introduction

Human translations are essential to reference-based MT evaluation metrics such as BLEU [10], NIST [7] and METEOR [1]. Generally there are a number of valid translations for a given source, varying in style, word order and word choice, but conveying similar meaning. As many as possible references substantially improve the reliability of n-gram based metrics as demonstrated by HyTER, which employs an exponential number of reference translations for a given target [8]. Although most metrics can take advantage of multiple references, the scheme is fairly simple by choosing the highest score out of pairwise comparison between the candidate and the reference. Using the highest score to estimate translation quality actually implies that the matching approach always underestimates the real quality of the candidate: the more matching, the closer estimation to the real quality. Nevertheless there is neither full argument nor empirical data to underpin this assumption.

Furthermore, since they originate from the same source segment, multiple references are expected to share words and expressions, but this kind of information is usually under-explored in reference-based metrics. For example:

Ref$_1$: *The report also shows that the US personal income rose 0.4% last December.*

Ref$_2$: *The report also indicated that U.S. personal income increased by 0.4 percent in December last year.*

Ref$_3$: *The report also shows that Americans' incomes rose by 0.4% last December.*

Ref$_4$: *The report also shows that the income of US individuals increased 0.4% last December.*

Sys: *The report also showed that in December last year the US personal income rose by 0.4 %.*

In the above, some words like *report, shows, income, 0.4%* appear frequently in references and convey the core meaning of the source sentence. However the recurring information is ignored by the current metrics. As a consequence, the score of this system translation according to these metrics can not reflect its quality (METEOR score for it is 0.403).

We suggest that common information in multiple references should appear in a good quality translation and therefore should be considered more relevant than non-recurring information by evaluation metrics. In our previous work, common information of multiple references is attempted to improve BLEU and NIST metrics [11]. As a further study, among a wide range of non-exact n-gram matching metrics we pick METEOR, which is reported to perform well in WMT evaluation task [2], and can be equally applied to segment and system-level evaluation. We thus propose a modification of METEOR to truly explore multiple references. First we provide an analysis comparing several approaches to handle scores produced by different references. Second we investigate how to make better use of common information in multiple references.

In the section follows we briefly review the METEOR metric. In Section 3 we discuss different ways of taking into account scores from multiple references and our proposal of two weighting strategies for METEOR to make better use of recurring information. Experiments with two into-English datasets are presented in Section 4.

2 METEOR Review

Based on word to word alignment of the candidate and the reference, METEOR calculates precision (P), recall (R) and final score (F). P is the ratio of the number of unigrams matched to the total number of unigrams in candidate. And R is computed as the number of unigrams matched divided by the total unigrams in reference. In addition to exact matching of words, METEOR allows for morphological variants and synonym matching to capture more similarities between candidate and references. A penalty is applied to favor the long chunks(sequential matches) over short ones. F_{mean}, the harmonic mean of precision and recall, combined with the penalty factor is used as the final score [1].

Recent improvements on METEOR include the matching of paraphrases for phrases [4], adjusting weights of content words versus function words, and a parameterized penalty [5]. The latest version of METEOR (v1.5) uses automatic extraction of language sources (including paraphrases and function words) and universal parameters in score formula to cope with specific translations [6]. The components F_{mean} and $Penalty$, as well as the final METEOR F_{score} are given in Equations 1-3 [6].

$$F_{mean} = \frac{PR}{\alpha P + (1 - \alpha)R} \tag{1}$$

$$Penalty = \gamma(\frac{ch}{m})^{\beta} \tag{2}$$

$$F_{score} = F_{mean} * (1 - Penalty) \tag{3}$$

where, α, γ and β are parameters tuned on corpora to maximize the correlation with human judgments. Equation 2 is used to favor longer matches by considering the number of chunks matched ch and the number of unigrams matched m.

When multiple references are available, METEOR picks the one with the highest score, i.e. the closest reference to the MT output. The intuition behind this choice is one of the questions we put forward in this paper.

3 Multiple References in METEOR

3.1 Handling Multiple Scores

In what follows we compare five options on the use of the final METEOR scores generated under the circumstance of multiple references: the highest score (current choice in METEOR), the lowest score and the arithmetic mean (AM), geometric mean (GM) and harmonic mean (HM) of all scores, as shown in Equations 4-6, where n is the number of references.

$$AM = \frac{1}{n}\sum_{i} x_i \tag{4}$$

$$GM = (\prod_{i} x_i)^{\frac{1}{n}} \tag{5}$$

$$HM = \frac{n}{\sum_i x_i^{-1}} \tag{6}$$

The highest score from multiple references is the one that results in the best quality score for the candidate. We believe that this leads to the quality of candidate being overestimated by METEOR.

3.2 Recurring Information in Multiple References

In the work of [6], several preferences are found in human judgments of translation quality:

- Precision is preferred over recall
- Correct word choice is preferred over correct word order
- Correct choice of content words is preferred over correct choice of function words
- Exact matching over matching of paraphrases

The repeated words in different references are expected relevant to correct word choice and content of translations which will in turn benefit accuracy evaluation. We thus propose two weighting strategies to make better use of the information and implement modification on standard METEOR metric.

The first weighting strategy is shown as Equation 7. We define the ratio of the number of times a word recurs in references to the number of references as the degree of commonality of the word. We add this ratio in a logarithm function as the weight of the matching word. In order to avoid zero counts, a plus-one smoothing approach is applied.

$$X = log(1 + m/refno) \tag{7}$$

In Equation 7, m denotes the number of times the word recurs. Notice that repeated words in single reference are not counted, so the ratio is never greater than 1. For the example above, the weight of *report*, which is covered by four references, is 0.69, heavier than *by* 0.41, which occurs twice.

Alternatively, we apply Zipf'law to lessen the impact of non-content words which may appear more often than content words, as defined in Equation 8.

$$X' = log(1 + f \times r/refno) \tag{8}$$

where f is the frequency of word in references and r is the ranking order of the word by frequency. For the example above, the weight of *the*, which has the most frequency, is 0.92, comparing with 1.18 of the third-ranking word *income*. The most frequent non-content words can be neutralized.

Precision and recall are updated with the weighting of matching words accordingly. Since we cannot estimate the weight of unmatched words in candidate and references, we keep them unchanged. Precision and recall are updated as in Equations 9-10.

$$P = \frac{\sum x_i w_i}{\#UnmatchedWordsInSys + \sum x_i w_i} \tag{9}$$

$$R = \frac{\sum x_i w_i}{\#UnmatchedWordsInRef + \sum x_i w_i} \tag{10}$$

where x_i is the weight of word w_i according to Equation 7 or 8. Obviously w_i is always 1 in normal METEOR. For simplicity, we use a fixed penalty as in earlier versions of METEOR. Accordingly, penalty is normalized by the sum of matching weights. Therefore the equations for F_{mean} and final score F remain the same as in standard METEOR.

4 Experiments

METEOR 0.4.3[1] (Perl version) is used in the experiment to test our two ideas. We also compare the performance against the latest version of the metrics, METEOR 1.5[2].

4.1 Datasets

The experiments require datasets with multiple references, which are rare. Two into-English translation datasets are used: the Multiple-Translation Chinese Part 2 (MTC-P2) (LDC2003T17) and the Multiple-Translation Chinese Part 4 (MTC-P4) (LDC2006T04), both including 4 sets of human translations for a single set of Mandarin Chinese source materials, totally 200 stories, 1797 segments. Altogether, the two datasets have nine system translations P2-05, P2-09, P2-14, P4-09, P4-11, P4-12, P4-14, P4-15 and P4-22, judged by 2-3 human annotators on fluency and accuracy. We use these judgments as ground truth to compare metrics. Notice that the Cohen's Kappa coefficient [3] of human judgments on segment level is only fair: 0.227 on fluency and 0.172 on accuracy annotation.

We investigate the variation among references in terms of TER (Translation Error Rate) [3] [12]. The average pairwise TER values of references for the two datasets are 0.72 and 0.67, indicating remarkable differences among the four references.

4.2 Multiple Score Selections

In order to examine the impact of synonym mapping, we run METEOR with two types of modules for alignment: exact matching and stemming module ($AS1$), versus added synonym matching ($AS2$).

The general relationship between the five score selections from references is as below.

$$HS \geq AMS \geq GMS \geq HMS \geq LS \tag{11}$$

where HS, AMS, GMS, HMS and LS denote the highest score, arithmetic mean, geometric mean, harmonic mean and the lowest score respectively.

Performances of METEOR based on different score selections on fluency and accuracy with and without synonym matching at system-level in terms of Pearson correlation are shown in Tables 1-2. The introduction of synonym alignment

[1] http://www.cs.cmu.edu/ banerjee/MT/METEOR/
[2] http://www.cs.cmu.edu/ alavie/METEOR/
[3] http://www.cs.umd.edu/ snover/tercom/

in METEOR significantly improves the correlation with human judgments on accuracy evaluation. However, it does not increase the performance on fluency evaluation.

Regardless of the type of alignment (with or without synonym matching) and the evaluation criteria (fluency or accuracy), the highest correlation with human judgment is always achieved when the reference with the *lowest* matching score is selected. Intuitively, the closer to the reference, the better quality the translation should have. Therefore, these results are somewhat puzzling and against the general practice in the use of METEOR with multiple references: that of choosing the closest matching reference.

Table 1. Correlation at system level with module $AS1$

		HS	AMS	GMS	HMS	LS
Flu	P	0.693	0.695	0.696	**0.699**	0.695
	R	0.354	0.367	0.371	0.376	**0.394**
	F	0.470	0.500	0.504	0.509	**0.545**
Acc	P	**0.741**	0.740	0.739	0.734	0.736
	R	0.863	0.869	0.870	0.872	**0.882**
	F	0.899	0.905	0.905	0.906	**0.913**

Table 2. Correlation at system level with module $AS2$

		HS	AMS	GMS	HMS	LS
Flu	P	0.729	0.729	0.730	0.731	**0.738**
	R	0.321	0.338	0.342	0.344	**0.365**
	F	0.447	0.486	0.490	0.494	**0.542**
Acc	P	**0.773**	0.766	0.765	0.764	0.758
	R	0.858	0.866	0.868	0.869	**0.878**
	F	0.909	0.919	0.920	0.921	**0.931**

In the remaining experiments, we use alignment module $AS2$, as the trend observed for score selection options with and without synonym matching is consistent.

In Tables 3-4 we compare the five score selection approaches at segment level. On fluency evaluation, the best score is obtained with the reference with the highest (HS) score in most cases. However, for accuracy evaluation, it seems better to choose the arithmetic mean of scores (AMS), as the arithmetic mean scores outperform in more than half system translations.

Table 3. Fluency evaluation correlation at segment level

System	HS	AMS	GMS	HMS	LS
p2-05	**0.213**	0.186	0.173	0.161	0.116
p2-09	**0.106**	0.100	0.085	0.083	0.075
p2-14	**0.218**	0.203	0.206	0.198	0.149
p4-09	**0.217**	0.200	0.194	0.190	0.173
p4-11	**0.164**	0.163	0.160	0.158	0.144
p4-12	**0.061**	0.049	0.046	0.041	0.017
p4-14	**0.155**	0.133	0.126	0.120	0.080
p4-15	0.196	**0.197**	0.190	0.185	0.162
p4-22	0.171	**0.172**	0.166	0.166	0.144

4.3 Recurring Word Weighting

The best correlations are usually achieved using either the highest or the lowest score out of multiple references. We thus focus on these score selection comparisons for the weighting strategies proposed here to improve METEOR.

System level evaluation using X and X' Figures 1 and 2 illustrate the performance of the modified METEOR after weighting the matching words by using recurring information in references at system level. The performance increases consistently when the weighting strategy is applied, regardless of whether the highest or the lowest scores are used (O denotes the original approach). The strategy using Zipf'law (X') is weaker than the alternative strategy proposed (X). The possible reason is that Zipf'law does not work well on small scale of corpus.

From v1.3 on, METEOR discriminates content words and function words [5] by weighting differently. All parameters of METEOR v1.5 run in the experiment are set as the default values for English translation evaluation. While in METEOR v0.4.3, all matching words are equally weighted. By using common information of multiple references, we modify METEOR v0.4.3 (denoted

Table 4. Accuracy evaluation correlation at segment level

System	HS	AMS	GMS	HMS	LS
p2-05	0.281	**0.284**	0.275	0.265	0.221
p2-09	0.143	**0.156**	0.149	0.146	0.123
p2-14	**0.245**	0.231	0.222	0.214	0.175
p4-09	**0.293**	0.279	0.276	0.272	0.247
p4-11	0.208	**0.216**	0.213	0.209	0.188
p4-12	**0.199**	0.182	0.174	0.166	0.125
p4-14	**0.194**	0.172	0.171	0.164	0.131
p4-15	0.250	**0.256**	0.251	0.246	0.219
p4-22	0.190	**0.202**	0.188	0.192	0.184

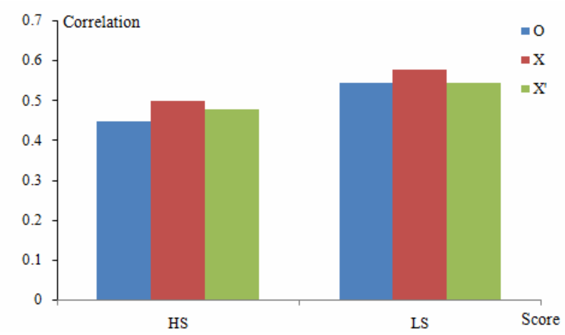

Fig. 1. Modified METEOR at system level on fluency

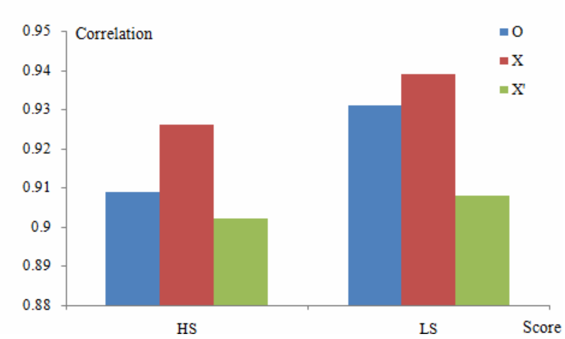

Fig. 2. Modified METEOR at system level on accuracy

as v0.4.3+) and compare it against the latest version v1.5 at system level using the highest score out of references.

This comparison is shown in Table 5. v0.4.3+ outperforms v1.5 in terms of accuracy, but it falls behind v1.5 in fluency evaluation. Nevertheless, the improvement with respect to v0.4.3 w.r.t fluency is evident. This confirms our intuition that taking recurring information of references into account is a sound strategy for MT evaluation.

Table 5. Comparison of METEOR versions at system level

	v1.5	v0.4.3	v0.4.3+
Flu	**0.619**	0.447	0.497
Acc	0.907	0.909	**0.926**

Segment level evaluation using X The proposed modification at segment level does not work well as at system level. Table 6 shows there is a slight drop of correlation with human judgment by introducing weighting strategy X (marked with +), as compared to the original METEOR on both fluency and accuracy evaluation.

Table 6. Segment level correlation comparison

System	Flu	Flu+	Acc	Acc+
p2-05	**0.213**	0.207	0.281	**0.284**
p2-09	**0.106**	0.102	**0.143**	0.140
p2-14	**0.218**	0.214	**0.245**	0.233
p4-09	**0.217**	0.204	**0.293**	0.279
p4-11	**0.164**	0.156	0.208	0.208
p4-12	**0.061**	0.051	**0.199**	0.179
p4-14	**0.155**	0.152	**0.194**	0.181
p4-15	**0.196**	0.179	**0.250**	0.113
p4-22	0.171	**0.176**	**0.190**	0.188

Due to the large TER values among the references, the data sparsity is severe at segment level. Therefore the most common words among references might not always be the content words especially for short sentences. For the example above, *would* is undesirably assigned more heavily than *American*. Data sparsity might be the main cause of performance drop at segment level. In addition, the low inter-agreements in human assessments pose challenges on the improvement of quality evaluation at segment level [9].

5 Conclusion and Future Work

We compared five score selection strategies for METEOR to handle multiple references and proposed to weight differently matching words by taking recurring information of these words in references into account. Results show that it is not always wise to select the reference with the highest matching score, especially at system level evaluation. It seemed we overestimated the translation quality by mean of alignment with candidate and references, contrary to the intuitive assumption in current reference-based evaluation metrics. Generally, the recurring information in references proved helpful to translation quality evaluation.

In future work, we will explore more common features of multiple references like POS and syntactic structures and integrate them into MT evaluation metrics.

Acknowledgments

Ying Qin's work is supported by a Beijing Social Science Funding Project (No. 15WYA006) and National Research Center for Foreign Language Education, BFSU.

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72 (2005)
2. Bojar, O., Buck, C., Federmann, C., et al.: Findings of the 2014 workshop on statistical machine translation. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. pp. 12–58 (2014)
3. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement. vol.20. pp. 37–46 (1960)
4. Denkowski, M., Lavie, A.: Extending the meteor machine translation evaluation metric to the phrase level. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 250–253. Association for Computational Linguistics (2010)
5. Denkowski, M., Lavie, A.: Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 85–91. Association for Computational Linguistics (2011)

6. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the EACL 2014 Workshop on Statistical Machine Translation. vol. 6 (2014)
7. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the second international conference on Human Language Technology Research. pp. 138–145 (2002)
8. Dreyer, M., Marcu, D.: Hyter: Meaning-equivalent semantics for translation evaluation. In: 2012 Conference of the North American Chapter of the ACL: Human Language Technologies. pp. 162–171 (2012)
9. Graham, Y., Mathur, N., Baldwin, T.: Accurate evaluation of segment-level machine translation metrics. In: Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL. pp. 1183–1191 (2015)
10. Papineni, K., Roukos, S., Ward, T., et al.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. ACL. pp. 311–318 (2002)
11. Qin Y., Specia, L.:Truly exploring multiple references for machine translation evaluation. In: Proceedings of European Association for machine translation (EAMT) pp. 113–120 (2015)
12. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of association for machine translation in the Americas. pp. 223–231 (2006)