

Word Choice and Word Position for Automatic MT Evaluation

Billy Tak-Ming Wong **Chunyu Kit**

Department of Chinese, Translation and Linguistics
City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
{ctbwong, ctckit}@cityu.edu.hk

1 Introduction

We describe our metric, ATEC¹, for automatic machine translation evaluation via two essential features of translation quality: word choice and word position. Its computation is based on unigram F-measure to rate the matching of words between candidate and reference translations, and average difference of relative position of the matched words.

2 The ATEC Metric

2.1 Unigram-based measure of word choice

We measure the word choice of a translation by unigram matching rate, which can be represented by the standard measures of precision (P) and recall (R), where the number of matched unigram (M) between a candidate translation (c) and a reference translation (r) is divided by the length of candidate translation ($|c|$) and reference translation ($|r|$) respectively:

$$P(c, r) = \frac{M(c, r)}{|c|} \quad R(c, r) = \frac{M(c, r)}{|r|}$$

As to rely on precision or recall in isolation may be over- or under-rate a candidate translation with less or more words than its reference translation, the use of their average F-measure (F) can avoid this:

$$F(c, r) = \frac{2P(c, r)R(c, r)}{P(c, r) + R(c, r)}$$

To avoid double-counting, every candidate word is considered exhausted once it matches a reference word. We further adopt the idea of METEOR

(Banerjee & Lavie, 2005) for the use of WordNet to enhance the matching of synonyms. After an exact matching is performed, a WordNet module tries to search for synonyms from the remaining words.

When multiple reference translations are available, we try to maximize the matches between a candidate and its references. As in example 1², the candidate shares 8 words with reference 1 (in underline) and 6 words with reference 2 (in italic) respectively. If we extend the matching to both references, there will be 9 matched words. We think that this is a better way to utilize reference translations as to provide more variation of possible word choices.

In the computation of recall under the situation of multiple references, the average reference length is used as the denominator, which also serves as an upper limit of the number of matches. Extra matches will not be counted. This avoids the case of a recall larger than 1.

Example 1:

Reference 1: US treasury offers 14 billion of 30
year treasury bonds

Reference 2: American *treasury* department
auctions *14 million dollars'* worth of *30* year
maturity *bonds*

Candidate: The US treasury offers 14 billion
dollars of bonds with a due term for 30 years

It may be the case that an MT system tries to game the metric by generating more synonyms in its outputs to match more words of multiple references, such as in example 2. This will be penalized by its longer length in precision, and a larger pe-

¹ ATEC: Assessment of Text Essential Characteristics

² Excerpted from the development data of MATR08 with some modifications.

nalty of word position difference which will be illustrated in next subsection.

Example 2:

Candidate: The American US treasury department offers auctions 14 billion dollars worth of treasury maturity bonds with a due term for 30 years

2.2 Penalty of word position difference

Instead of strictly requiring those matched words in consecutive order as n-gram based metrics such as BLEU (Papineni et al., 2001) does, we propose another measurement based on word position. We think that every word has its proper word position in a sentence to contribute to a particular sentence meaning. For instance, in example 3a, candidate 1 has a different meaning from the reference although they share the same words. Candidate 2 shares many consecutive words with the reference (i.e., “chases a thief”, “the police”), but it is ungrammatical. Candidate 3 shares the least words and none of bigrams or above with the reference, but it has the closest meaning to it.

Example 3a:

Reference: the police chase the thief
Candidate 1: the thief chase the police
Candidate 2: chase the thief the police
Candidate 3: a police quickly chase a thief

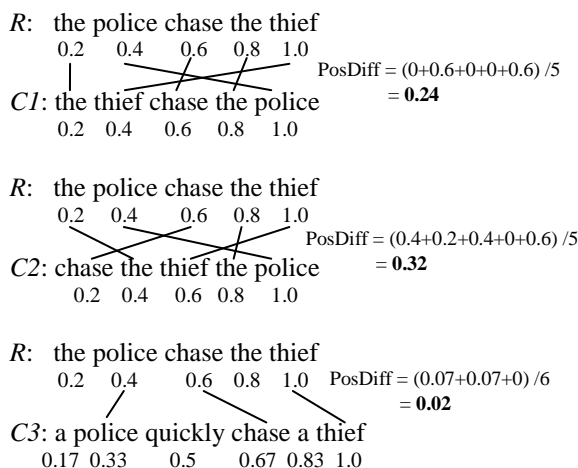
To account for their variances in word order, we first assign an absolute position to each of the words of both candidates and references. The absolute positions are then converted to relative positions by dividing them to the lengths of candidate or reference in order to normalize the length difference of each sentence, as shown in example 3b.

Example 3b:

Reference: the police chase the thief
 Absolute position: 1 2 3 4 5
 Relative position: 0.2 0.4 0.6 0.8 1.0
Candidate 1: the thief chase the police
 Absolute position: 1 2 3 4 5
 Relative position: 0.2 0.4 0.6 0.8 1.0
Candidate 2: chase the thief the police
 Absolute position: 1 2 3 4 5
 Relative position: 0.2 0.4 0.6 0.8 1.0
Candidate 3: a police quickly chase a thief
 Absolute position: 1 2 3 4 5 6
 Relative position: 0.17 0.33 0.5 0.67 0.83 1.0

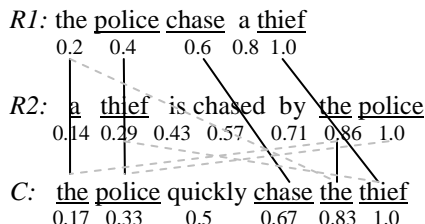
For each candidate, their words are aligned to corresponding words in the reference. If a candidate word can be aligned to more than one reference words, like the word “the” in *C1* and *C2* of example 3c, the reference word in its closest relative position is chosen. After this alignment process, a word position difference between a candidate and a reference is then calculated by dividing the summation of the differences of relative position of each candidate-reference word alignment with candidate sentence length, like the calculations in example 3c.

Example 3c:



In the case of multiple references, each candidate word is aligned to a corresponding reference word of any references in its closest relative position. As shown in example 3d, the candidate words “the”, “police” and “thief” are aligned to their closest counterparts in *R1* (in black lines) rather than in *R2* (in grey lines).

Example 3d:



After a word position difference between a candidate and one or more references is calculated, it is then converted to a penalty rate for the candidate. According to our empirical experiment, the

word position difference has to be multiplied by a coefficient 4 as follows for the highest correlation with human judgment.

$$Penalty = 1 - (PosDiff * 4)$$

If the word position difference (PosDiff) of a candidate is larger than 0.25, like *C2* of example 3c, the penalty will be smaller than 0. In this case the penalty will be adjusted to 0.

2.3 Formulation of ATEC

Currently the computation of ATEC is on sentence level. By default, punctuations are removed from each candidate and reference translation, and the measurement is performed in a case insensitive manner. For each candidate sentence, the ATEC score is computed by combining the unigram F-measure and the penalty of word position difference in the following formula:

$$ATEC = F(c, r) * Penalty$$

After the ATEC scores of each candidate sentences of a system are calculated, they are averaged into a corpus level score for that system.

3 Conclusion

We have presented an MT evaluation metric, ATEC, which tries to model two essential features of translation quality, i.e., word choice and word position. We try to provide independent measures of them such that each feature can be manipulated and optimized in a flexible way. In this way, the resultant score is intuitive. We can easily relate between unigram F-measure and word choice, and between average word position difference and word position; hence understand the performance of a system in more details apart from merely a system ranking.

The current ATEC metric only encodes two features of translation quality. We are sure that there are many others that can be included. In our viewpoint, a holistic perspective of MT evaluation based on multiple facets of translation quality is the final solution of MT evaluation. Now we already have many useful evaluation metrics operating on the words shared by candidate and reference translations. Our current focus is on the unmatched

words which is still an explored field waiting for harvest. Besides, we will continue the study of the performance of ATEC in different languages as well as in the evaluation of other language technologies.

For further information about ATEC and its implemented version, please visit the following site:

<http://144.214.20.216:8080/ctbwong/ATEC/index.html>

References

- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022).
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*. Ann Arbor, Michigan, June 2005.