The Parameter-optimized ATEC Metric for MT Evaluation

Billy T-M 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

Abstract

This paper describes the latest version of the ATEC metric for automatic MT evaluation, with parameters optimized for word choice and word order, the two fundamental features of language that the metric relies on. The former is assessed by matching at various linguistic levels and weighting the informativeness of both matched and unmatched words. The latter is quantified in term of word position and information flow. We also discuss those aspects of language not yet covered by other existing evaluation metrics but carefully considered in the formulation of our metric.

1 Introduction

It is recognized that the proposal of the BLEU metric (Papineni et al., 2002) has piloted a paradigm evolution to MT evaluation. It provides a computable solution to the task and turns it into an engineering problem of measuring text similarity and simulating human judgments of translation quality. Related studies in recent years have extensively revealed more essential characteristics of BLEU, including its strengths and weaknesses. This has aroused the proposal of different new evaluation metrics aimed at addressing such weaknesses so as to find some other hopefully better alternatives for the task. Effort in this direction brings up some advanced metrics such as METEOR (Banerjee and Lavie, 2005) and TERp (Snover et al., 2009) that seem to have already achieved considerably strong correlations with human judgments. Nevertheless, few metrics have really nurtured our understanding of possible parameters involved in our language comprehension and text quality judgment. This inadequacy limits, inevitably, the application of the existing metrics.

The ATEC metric (Wong and Kit, 2008) was developed as a response to this inadequacy, with a focus to account for the process of human comprehension of sentences via two fundamental features of text, namely word choice and word order. It integrates various explicit measures for these two features in order to provide an intuitive and informative evaluation result. Its previous version (Wong and Kit, 2009b) has already illustrated a highly comparable performance to the few state-of-the-art evaluation metrics, showing a great improvement over its initial version for participation in MetricsMATR08[1]. It is also applied to evaluate online MT systems for legal translation, to examine its applicability for lay users' use to select appropriate MT systems (Wong and Kit, 2009a).

In this paper we describe the formulation of ATEC, including its new features and optimization of parameters. In particular we will discuss how the design of this metric can complement the inadequacies of other metrics in terms of its treatment of word choice and word order and its utilization of multiple references in the evaluation process.

2 The ATEC Metric

2.1 Word Choice

In general, word is the basic meaning bearing unit of language. In a semantic theory such as Latent Semantic Analysis (LSA) (Landauer et al., 1998), lexical selection is even the sole consideration of the meaning of a text. A recent study of the major errors in MT outputs by Vilar et al. (2006) also reveals that different kinds of error related to word choices constitute a majority of error types. It is therefore of prime importance

[1] http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2008/

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 360–364,
Uppsala, Sweden, 11-16 July 2010.

for MT evaluation metrics to diagnose the adequacy of word selection by an MT system.

It is a general consensus that the performance of an evaluation metric can be improved by matching more words between MT outputs and human references. Linguistic resources like stemmer and WordNet are widely applied by many metrics for matching word stems and synonyms. ATEC is equipped with these two modules as well, and furthermore, with two measures for word similarity, including a WordNet-based (Wu and Palmer, 1994) and a corpus-based measure (Landauer et al., 1998) for matching word pairs of similar meanings. Our previous work (Wong, 2010) shows that the inclusion of semantically similar words results in a positive correlation gain comparable to the use of WordNet for synonym identification.

In addition to increasing the number of legitimate matches, we also consider the importance of each match. Although most metrics score every matched word with equal weight, different words indeed contribute different amount of information to the meaning of a sentence. In Example 1 below, both *C1* and *C2* contain the same number of words matched with *Ref*, but the matches in *C1* are more informative and therefore should be assigned higher weights.

Example 1
C1: it was not first time that prime minister confronts northern league …
C2: this is not the prime the operation with the north …
Ref: this is not the first time the prime minister has faced the northern league …

The informativeness of a match is weighted by the *tf-idf* measure, which has been widely used in information retrieval to assess the relative importance of a word as an indexing term for a document. A word is more important to a document when it occurs more frequently in this document and less in others. In ATEC, we have "document" to refer to "sentence", the basic text unit in MT evaluation. This allows a more sensitive measure for words in different sentences, and gets around the problem of an evaluation dataset containing only one or a few long documents. Accordingly, the *tf-idf* measure is formulated as:

$$tfidf(i,j) = tf_{i,j} \cdot \log(\frac{N}{sf_i})$$

where $tf_{i,j}$ is the occurrences of word w_i in sentence s_j, sf_i the number of sentences containing word w_i, and N the total number of sentences in the evaluation set. In case of a high-frequency word whose *tf-idf* weight is less than 1, it is then rounded up to 1.

In addition to matched words, unmatched words are also considered to have a role to play in determining the quality of word choices of an MT output. As illustrated in Example 1, the unmatched words in *Ref* for *C1* and *C2* are [this | is | the | the | has | faced | the] and [first | time | minister | has | faced | northern | league] respectively. One can see that the words missing in *C2* are more significant. It is therefore necessary to apply the *tf-idf* weighting to unmatched reference words so as to quantify the information missed in the MT outputs in question.

2.2 Word Order

In MT evaluation, word order refers to the extent to which an MT output is interpretable following the information flow of its reference translation. It is not rare that an MT output has many matched words but does not make sense because of a problematic word order. Currently it is observed that consecutive matches represent a legitimate local ordering, causing some metrics to extend the unit of matching from word to phrase. Birch et al. (2010) show, however, that the current metrics including BLEU, METEOR and TER are highly lexical oriented and still cannot distinguish between sentences of different word orders. This is a serious problem in MT evaluation, for many MT systems have become capable of generating more and more suitable words in translations, resulting in that the quality difference of their outputs lies more and more crucially in the variances of word order.

ATEC uses three explicit features for word order, namely position distance, order distance and phrase size. Position distance refers to the divergence of the locations of matches in an MT output and its reference. Example 2 illustrates two candidates with the same match, whose position in *C1* is closer to its corresponding position in *Ref* than that in *C2*. We conceive this as a significant indicator of the accuracy of word order: the closer the positions of a matched word in the candidate and reference, the better match it is.

Example 2
C1: non-signatories these acts victims but it caused to incursion transcendant
C2: non-signatories but it caused to incursion transcendant these acts victims
Ref: there were no victims in this incident but they did cause massive damage

The calculation of position distance is based on the position indices of words in a sentence. In particular, we align every word in a candidate to its closest counterpart in a reference. In Example 3, all the candidate words have a match in the reference. As illustrated by the two "*a*" in the candidate, the shortest alignments (strict lines) are preferred over any farther alternatives (dash lines). In a case like this, only two matches, i.e., *thief* and *police*, vary in position by a distance of 3.

Example 3

Candidate:	a	thief	chases	a	police	
Pos distance:	0	3	0	0	3	
Pos index:	1	2	3	4	5	
Reference:	a	police	chases	a	thief	
Pos index:	1	2	3	4	5	

This position distance is sensitive to sentence length as it simply makes use of word position indices without any normalization. Example 4 illustrates two cases of different lengths. The position distance of the bold matched words is 3 in *C1* but 14 in *C2*. Indeed, the divergence of word order in *C1* does not hinder our understanding, but in *C2* it poses a serious problem. This excessive length inevitably magnifies the interference effect of word order divergence.

Example 4

C1: Short$_1$ and$_2$ various$_3$ **international$_4$** news$_5$

R1: **International$_1$** news$_2$ brief$_3$

C2: Is$_1$ on$_2$ a$_3$ popular$_4$ the$_5$ very$_6$ in$_7$ Iraq$_8$ to$_9$ those$_{10}$ just$_{11}$ like$_{12}$ other$_{13}$ world$_{14}$ in$_{15}$ which$_{16}$ young$_{17}$ people$_{18}$ with$_{19}$ the$_{20}$ and$_{21}$ flowers$_{22}$ while$_{23}$ awareness$_{24}$ by$_{25}$ other$_{26}$ times$_{27}$ of$_{28}$ the$_{29}$ **countries$_{30}$** of$_{31}$ the$_{32}$

R2: Valentine's$_1$ day$_2$ is$_3$ a$_4$ very$_5$ popular$_6$ day$_7$ in$_8$ Iraq$_9$ as$_{10}$ it$_{11}$ is$_{12}$ in$_{13}$ the$_{14}$ other$_{15}$ **countries$_{16}$** of$_{17}$ the$_{18}$ world$_{19}$. Young$_{20}$ men$_{21}$ exchange$_{22}$ with$_{23}$ their$_{24}$ girlfriends$_{25}$ sweets$_{26}$, flowers$_{27}$, perfumes$_{28}$ and$_{29}$ other$_{30}$ gifts$_{31}$.

Another feature, the order distance, concerns the information flow of a sentence in the form of the sequence of matches. Each match in a candidate and a reference is first assigned an order index in a sequential manner. Then, the difference of two counterpart indices is measured, so as to see if a variance exists. Examples 5a and 5b exemplify two kinds of order distance and their corresponding position distance. Both cases have two matches with the same sum of position distance. However, the matches are in an identical sequence in 5a but cause a cross in 5b, resulting in a larger order distance for the latter.

Example 5a

Position index	1	2	3	4
Order index		1		2
Candidate:	A	B	C	D
Reference:	B	E	D	F
Order index	1		2	
Position index	1	2	3	4

Position distance (2-1) + (4-3) = 2

Order distance (1-1) + (2-2) = 0

Example 5b

Position index	1	2	3	4
Order index		1	2	
Candidate:	A	B	C	D
Reference:	C	B	E	F
Order index	1	2		
Position index	1	2	3	4

Position distance (2-2) + (3-1) = 2

Order distance (2-1) + (2-1) = 2

In practice, ATEC operates on phrases like many other metrics. But unlike these metrics that count only the number of matched phrases, ATEC gives extra credit to a longer phrase to reward its valid word sequence. In Example 6, *C1* and *C2* represent two MT outputs of the same length, with matched words underlined. Both have 10 matches in 3 phrases, and will receive the same evaluation score from a metric like METEOR or TERp, ignoring the subtle difference in the sizes of the matched phrases, which are [8,1,1] and [4,3,3] words for *C1* and *C2* respectively. In contrast, ATEC uses the size of a phrase as a reduction factor to its position distance, so as to raise the contribution of a larger phrase to the metric score.

Example 6

C1: $\underline{w_1}$ $\underline{w_2}$ $\underline{w_3}$ $\underline{w_4}$ $\underline{w_5}$ $\underline{w_6}$ $\underline{w_7}$ $\underline{w_8}$ w_9 w_{10} $\underline{w_{11}}$ w_{12} $\underline{w_{13}}$

C2: $\underline{w_1}$ $\underline{w_2}$ $\underline{w_3}$ $\underline{w_4}$ w_5 w_6 $\underline{w_7}$ $\underline{w_8}$ $\underline{w_9}$ w_{10} $\underline{w_{11}}$ $\underline{w_{12}}$ $\underline{w_{13}}$

2.3 Multiple References

The availability of multiple references allows more legitimate word choices and word order of an MT output to be accounted. Some existing metrics only compute the scores of a candidate against each reference and select the highest one.

This deficit can be illustrated by a well-known example from Papineni et al. (2002), as replicated in Example 7 with slight modification. It shows that nearly all candidate words can find their matches in either reference. However, if we resort to single reference, only around half of them can have a match, which would seriously underrate the quality of the candidate.

Example 7

C: It is a guide to action which ensures that the military always obeys the commands of the party.

R1: It is a guide to action that ensures that the military will forever heed Party commands.

R2: It is the guiding principle which guarantees the military forces always being under the commands of the party.

ATEC exploits multiple references in this fashion to maximize the number of matches in a candidate. It begins with aligning the longest matches with either reference. The one with the shortest position distance is preferred if more than one alternative available in the same phrase size. This process repeats until no more candidate word can find a match.

2.4 Formulation of ATEC

The computation of an ATEC score begins with alignment of phrases, as described above. For each matched phase, we first sum up the score of each word *i* in the phrase as

$$W_{match} = \sum_{i \in \{phrase\}} (w_{type} - \frac{Info_{match}}{tfidf_i})$$

where w_{type} refers to a basic score of a matched word depending on its match type. It is then minus its information load, i.e., the *tf-idf* score of the matched word with a weight factor, $Info_{match}$.

There is also a distance penalty for a phrase,

$$Dis = w_{pos} dis_{pos} (1 - \frac{|p|^e}{|c|}) + w_{order} dis_{order},$$

where dis_{pos} and dis_{order} refer to the position distance and order distance, and w_{pos} and w_{order} are their corresponding weight factors, respectively. The position distance is further weighted according to the size of phrase $|p|$ with an exponential factor *e*, in proportion to the length of candidate $|c|$.

The score of a matched phrase is then computed by

$$Phrase = \begin{cases} W_{match} \cdot Limit_{dis}, & \text{if } Dis > W_{match} \cdot Limit_{dis}; \\ W_{match} - Dis, & \text{otherwise,} \end{cases}$$

$Limit_{dis}$ is an upper limit for the distance penalty. Accordingly, the score *C* of all phrases in a candidate is

$$C = \sum_{j \in \{candidate\}} Phrase_j .$$

Then, we move on to calculating the information load of unmatched reference words $W_{unmatch}$, approximated as

$$W_{unmatch} = \sum_{k \in \{unmatch\}} (w_{type} - \frac{Info_{unmatch}}{tfidf_k}) .$$

The overall score *M* accounting for both the matched and unmatched is defined as

$$M = \begin{cases} C \cdot Limit_{Info}, & \text{if } W_{unmatch} > C \cdot Limit_{Info}; \\ C - W_{unmatch}, & \text{otherwise,} \end{cases}$$

$Limit_{Info}$ is an upper limit for the information penalty of the unmatched words.

Finally, the ATEC score is computed using the conventional *F*-measure in terms of precision *P* and recall *R* as

$$ATEC = \frac{PR}{\alpha P + (1-\alpha) R}$$

where $P = \frac{M}{|c|}$, $R = \frac{M}{|r|}$.

The parameter α adjusts the weights of *P* and *R*, and $|c|$ and $|r|$ refer to the length of candidate and reference, respectively. In the case of multiple references, $|r|$ refers to the average length of references.

We have derived the optimized values for the parameters involved in ATEC calculation using the development data of NIST MetricsMATR10 with adequacy assessments by a simple hill climbing approach. The optimal parameter setting is presented in Table 1 below.

3 Conclusion

In the above sections we have presented the latest version of our ATEC metric with particular emphasis on word choice and word order as two fundamental features of language. Each of these features contains multiple parameters intended to

Parameters	Values	
w_{type}	1	(exact match),
	0.95	(stem / synonym / semantically close),
	0.15	(unmatch)
$Info_{match}$	0.34	
$Info_{unmatch}$	0.26	
w_{pos}	0.02	
w_{order}	0.15	
e	1.1	
$Limit_{dis}$	0.95	
$Limit_{Info}$	0.5	
α	0.5	

Table 1 Optimal parameter values for ATEC

have a comprehensive coverage of different textual factors involved in our interpretation of a sentence. The optimal offsetting for the parameters is expected to report an empirical observation of the relative merits of each factor in adequacy assessment. We are currently exploring their relation with the errors of MT outputs, to examine the potential of automatic error analysis. The ATEC package is obtainable at: http://mega.ctl.cityu.edu.hk/ctbwong/ATEC/

Acknowledgments

The research work described in this paper was supported by City University of Hong Kong through the Strategic Research Grant (SRG) 7002267.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 65-72, Ann Arbor, Michigan, June 2005.

Alexandra Birch, Miles Osborne and Phil Blunsom. 2010. Metrics for MT Evaluation: Evaluating Reordering. *Machine Translation* (forthcoming).

Thomas Landauer, Peter W. Foltz and Darrell Laham. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes* 25: 259–284.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July 2002.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. *Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 259-268, Athens, Greece, March, 2009.

David Vilar, Jia Xu, Luis Fernando D'Haro and Hermann Ney. 2006. Error Analysis of Statistical Machine Translation Output. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 697-702, Genova, Italy, May 2006.

Billy T-M Wong. 2010. Semantic Evaluation of Machine Translation. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May, 2010.

Billy T-M Wong and Chunyu Kit. 2008. Word choice and Word Position for Automatic MT Evaluation. *AMTA 2008 Workshop: MetricsMATR*, 3 pages, Waikiki, Hawai'i, October, 2008.

Billy T-M Wong and Chunyu Kit. 2009a. Meta-Evaluation of Machine Translation on Legal Texts. *Proceedings of the 22nd International Conference on the Computer Processing of Oriental Languages (ICCPOL)*, pages 343-350, Hong Kong, March, 2009.

Billy Wong and Chunyu Kit. 2009b. ATEC: Automatic Evaluation of Machine Translation via Word Choice and Word Order. *Machine Translation*, 23(2):141-155.

Zhibiao Wu and Martha Palmer. 1994. Verb Semantics and Lexical Selection. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133-138, Las Cruces, New Mexico.