Mach Translat (2009) 23:141–155
DOI 10.1007/s10590-009-9061-x

ATEC: automatic evaluation of machine translation via word choice and word order

Billy Wong · Chunyu Kit

Received: 15 May 2009 / Accepted: 29 October 2009 / Published online: 12 December 2009

Abstract We propose a novel metric ATEC for automatic MT evaluation based on explicit assessment of word choice and word order in an MT output in comparison to its reference translation(s), the two most fundamental factors in the construction of meaning for a sentence. The former is assessed by matching word forms at various linguistic levels, including surface form, stem, sound and sense, and further by weighing the informativeness of each word. The latter is quantified in term of the discordance of word position and word sequence between a translation candidate and its reference. In the evaluations using the MetricsMATR08 data set and the LDC MTC2 and MTC4 corpora, ATEC demonstrates an impressive positive correlation to human judgments at the segment level, highly comparable to the few state-of-the-art evaluation metrics.

Keywords MT evaluation · Evaluation metrics · ATEC · Word choice · Word order

1 Introduction

Evaluation is one of the central concerns in machine translation (MT), not only because of its importance but also its difficulty, in that a satisfactory evaluation method is yet to be found. Human evaluation, following the general criteria for translation quality, such as adequacy and fluency, is highly susceptible to evaluation settings and other uncontrollable factors. Therefore, a large number of human judgments are needed in order to maintain its validity and reliability. There is a consensus that MT evaluation by humans is unrealistic in general, for it takes too much effort, cost and time (White 2000).

B. Wong (✉) · C. Kit
Department of Chinese, Translation and Linguistics, City University of Hong Kong,
83, Tat Chee Avenue, Kowloon, Hong Kong
e-mail: ctbwong@cityu.edu.hk

C. Kit
e-mail: ctckit@cityu.edu.hk

This difficult situation has changed since the pioneering work of BLEU scoring to enable automatic evaluation for MT (Papineni et al. 2002). This work provides not only an automatic cost-effective method to evaluate MT systems but also a rationale that the human judgment of the quality of an MT output can be quantified and simulated by measuring the textual similarity between the MT output and the available human translation(s) for the same source text. Such practice has inspired more efforts to investigate a variety of textual characteristics that can help emulate human judgment of the quality of MT output. Subsequently, a number of MT evaluation metrics were proposed to exploit different features of a text, ranging from the literal level, e.g., NIST (Doddington 2002), METEOR (Banerjee and Lavie 2005) and TER (Snover et al. 2006), to the syntactic level, e.g., STM and HWCM (Liu and Gildea 2005), and dependencies using Lexical-Functional Grammar (Owczarzak et al. 2007), and the semantic level, e.g., semantic role overlap (Giménez and Màrquez 2007).

In fact, all these textual characteristics are exploited to deal with the two basic issues in MT evaluation, namely, word choice and word order, which are two fundamental features used to determine the meaning of a sentence. In particular, they contribute critically to the adequacy and the fluency of a translation. It is then an interesting question how to formulate an appropriate measure to account for them.

In this paper we present our recent work on MT evaluation with a novel metric, namely ATEC,[1] that explicitly measures the word choice and word order in an MT output. It is proposed to incorporate the most critical textual features, in a way that can be verified through experiments that it can evaluate the performance of an MT system objectively and reliably. In the following sections we will first review the roles of word choice and word order in MT evaluation, and then present the details of the ATEC metric, including its rationale, its baseline version as submitted to the National Institute of Standards and Technology (NIST) MetricsMATR08, and the subsequent improvements. An empirical evaluation of the metric will be presented in Sect. 4, followed by a concluding section afterward.

2 Word choice and word order in MT evaluation

It is a complex issue to determine the exact contribution of word choice and word order to the meaning of a sentence in a language. For English, Landauer (2002) estimates that word choice constitutes about 80% of the basic information content of a text, while word order and other stylistic features account for the remaining 20%. At the sentence level, however, Gopen (2004) suggests that word choice only accounts for about 15% of the meaning and word order for the remaining 85%. Although this proportion varies in different studies, it is undoubted that word choice and word order together share the most critical contribution to the construction of meaning.

The main purpose of MT evaluation is to determine "to what extent the makers of a system have succeeded in mimicking the human translator" (Krauwer 1993). In an automatic manner, this is achieved by assessing the extent to which a system output simulates its human-translated version, with the aid of an evaluation metric. In

[1] ATEC: Assessment of Text Essential Characteristics.

principle, human evaluation follows a similar assessment process. As early as a half century ago Miller and Beebe-Center (1956) explained that "the fact that a grader can recognize errors at all implies that he must have some personal standard against which he compares the student's work… this might consist of his own written translation; more often it is probably a rather vague set of translations that would be about equally acceptable." They further proposed a primitive measure for "the relation between the test translation and the criteria", which is "to ask if they use the same words" and "to compare the order of the words which were common to the test and the criterion translations". It is an early idea for explicit assessment of translation in terms of word choice and word order. But it was not popularized until the BLEU metric.

The recent trend of research in automatic MT evaluation has resulted in many innovative and sophisticated measures for comparing word choice and word order of a system output and its reference translation. For the metrics relying on higher order n-grams, such as BLEU and NIST, word choice and word order are evaluated together, with certain rewards to the consecutive sequences of correctly matched words. The relative importance of a word sequence is rated by NIST in terms of its frequency in the dataset in question. Many other metrics deal with each word individually. For instance, METEOR uses unigram matching, in a way to allow words of the same stems and senses to match, by exploiting available NLP utilities and resources such as a stemmer and WordNet. The matched words not in adjacent positions are subject to a fragmentation penalty. TER and other edit-distance based metrics work in a different way to simulate the process of revising a system output into a humanlike version (i.e. the reference translation) by allowable editing actions on words such as addition, deletion and shift. Some metrics resort to linguistic features such as dependency and semantic role, looking for a theoretically sound manner to account for the legitimacy of the composition of words. All these measures have broadened our understanding of many possible parameters constituting our judgments of translation quality.

3 The ATEC metric

3.1 Rationale

The accuracy of word choice and word order of an MT output are not adequately assessed by the existing measures. Except NIST, which has a measure of information weight for matched n-grams, all others equally weight the matched words between a candidate and a reference translation regardless of their difference of importance in a sentence. As different words contribute a different amount of information to the meaning of a sentence, they should be weighted differently. This can be illustrated with the following examples.[2] In Example 1, it is reasonable to assign a higher score to Candidate 1 than Candidate 2 because the matched words in the former are more informative, even though the latter has more matched words.

[2] Selected from NIST MetricsMATR08 development data (LDC2009T05), with some modifications for illustration purpose.

Example 1

Candidate 1: it was not a case that prime minister confronts northern league … (5 matches)
Candidate 2: this is not the prime the cooperation with the north … (7 matches)
Reference: this is not the first time the prime minister has faced the northern league …

Besides rewarding consecutive matched words in terms of higher order n-grams or other means, there should also be a reasonable penalty for diverging word positions of the matched words between candidate and reference translations. In Example 2, both candidates have only one matched word, but its position in Candidate 1 is nearer to the corresponding position in the reference than that in Candidate 2. This can be a significant indicator of the accuracy of word order: the closer the positions of a matched word in the candidate and reference translation, the better match it is.

Example 2

Candidate 1: and non-signatories these acts victims but it caused to incursion transcendant
Candidate 2: and non-signatories but it caused to incursion transcendant these acts victims
Reference: there were no victims in this incident but they did cause massive damage

To address such inadequacies in MT evaluation, we attempt to integrate necessary measurement of word informativeness and word position distance into the ATEC metric. In general, it is formulated to compare a candidate against its reference in three aspects: the number of matched words, the importance of the un-/matched words, and their word order divergence. In the next subsection we will present the initial ATEC metric as submitted for our participation in NIST MetricsMATR08. A further elaboration of this metric will be presented in Sect. 3.3, followed by an illustration of its formula in detail in Sect. 3.4.

3.2 Original version

ATEC relies on unigram matching as a basis to compare word choice between a candidate translation and its reference. It is measured by the conventional precision P, recall R and F-measure defined as follows,

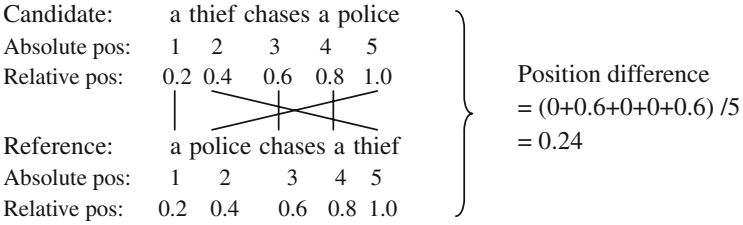
$$P = \frac{|M(c,r)|}{|c|} \qquad R = \frac{|M(c,r)|}{|r|} \qquad F = \frac{2PR}{P+R}$$

where $M(c,r)$ denotes the set of matched unigrams in a candidate translation c and its reference r, and in the denominator $|c|$ and $|r|$ are the number of words in each, respectively.

We further follow the idea of METEOR to use the WordNet to match synonyms. Once an exact matching is done, a WordNet module is applied to search for synonyms from the remaining unmatched words.

To account for the variances of word order between a candidate and its reference, the position differences of their matched words need to be measured. To do this, an absolute position is first assigned to each word, and then converted to a relative position with regard to the sentence length in question so as to normalize the length difference between the candidate and the reference. This kind of normalization is expected to allow a fair comparison of word positions in sentences of different lengths.

Example 3



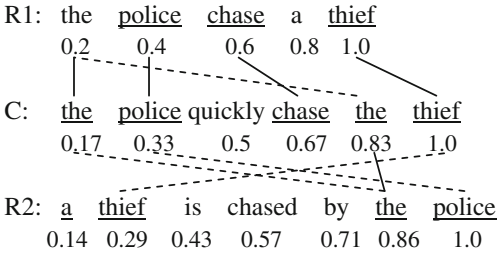
The relative word position is used in two ways. First, to match words between a candidate and a reference, if a candidate word can be aligned to more than one reference word, such as the word "a" in Example 3, the closest words in terms of their relative positions are matched. Second, the word position difference between a candidate and a reference sentence is calculated as the average relative position difference of all matched words in the candidate, as illustrated in Example 3. It is then converted to a penalty score to reflect the variance of word order using the following formula:

$$\text{Penalty}(c, r) = 1 - w \times d(c, r)$$

where $d(c, r)$ is the position difference and w is a coefficient to adjust the magnitude of the penalty. According to our experiments, setting w to four achieves the highest correlation to human judgment.

In the case of multiple references, each candidate word is aligned to a closest reference word in a reference. As illustrated in Example 4, the candidate words "police" and "thief" are aligned to their closest counterparts in R1 (in solid lines) rather than in R2 (in dotted lines). The average length of all references is used as the denominator for the calculation of recall, and as an upper limit for the number of possible matches. Extra matches will not be counted.

Example 4



Finally, the ATEC score of each candidate sentence is calculated as the product of unigram F-measure and the penalty for word position difference, as follows.

$$\text{ATEC}(c, r) = F(c, r) \times \text{Penalty}(c, r)$$

The ATEC score for a system is the average ATEC score over all its output sentences in question. By default, the word matching is case insensitive and punctuations are ignored.

3.3 Improvement

Based on the initial version as formulated above, ATEC is further elaborated by integrating new evaluation features. The major ones include matching word stems and homophones, calculating word informativeness, and extending the measure of word position distance, as described below.

Originally ATEC follows a two-stage matching strategy, i.e., the exact match first and then the synonym match, for which lemmatization is also applied with the aid of WordNet. But it still misses many legitimate matches, mostly those words (1) not in WordNet, (2) of multiple part-of-speeches, or (3) in different surface forms though phonetically similar or identical (especially various versions of transliteration such as *Nicolas*, *Nicolaas*, *Nicholas* and *Nicola*). To alleviate this problem, two available facilities, namely, the Porter stemming (Porter 1980) and the Soundex algorithm (Russell 1918) are applied. The former identifies word stems, hence allows words of the same stem to match with each other, regardless of their variances in inflections and derivations. The latter generates Soundex codes for words representing their pronunciations. Words having the same codes are considered phonetically identical, and hence allowed to match with each other. However, this also poses another problem that words of the same pronunciation are not necessarily related, particularly those high frequency words such as *this* versus *these* and *two* versus *to*. This problem has to be resolved heuristically according to word frequency by applying a constraint to the Soundex algorithm that the word pairs within the top 25% most frequent words are skipped. Consequently the matching process is extended from two- to four-stages: exact match, stem match, sound match and synonym match, so as to allow more legitimate matches at the levels of surface form, morphology, phonetics and semantics.

In addition to these flexible matches, the importance of each match is further measured in terms of their informativeness. This is performed with the aid of the conventional *tf-idf* measure in information retrieval, which assesses the relative importance of a word as an indexing term for a document. The more frequently a word occurs in a document and less in others, the more informative it is about this document. To integrate *tf-idf* into ATEC, however, we let a "document" refer to a "sentence", which is the basic unit in all current exercises of MT evaluation. This allows a *tf-idf* score to be more sensible in reflecting the information load of a word in a sentence, and avoids the risky use of the *tf-idf* measure in a situation of having an evaluation dataset of only a few long documents. It also distinguishes our approach from other similar works such as Babych and Hartley (2004). Accordingly, the *tf-idf* scoring is revised as:

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

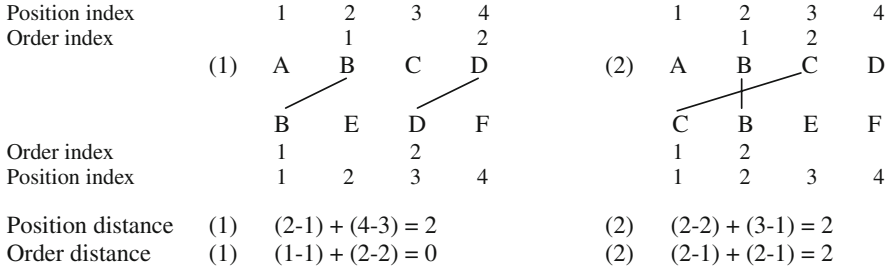
where $tf_{i,j}$ is the number of occurrences of word w_i in sentence s_j, sf_i the number of sentences containing word w_i, and N the total number of sentences in the dataset in question.

For those high frequency words with a *tf-idf* score <1.0, the score is rounded up to 1. The *tf-idf* measure is used to weight the informativeness of both matched and unmatched words. Since human evaluators tend to assign lower scores to MT outputs with many important words missing, it is thus reasonable that an assessment of the information load on unmatched words results in a higher correlation with human rating.

Several major extensions are made to the measurement of word position distance. Originally this distance is applied to a matched word as a penalty. Unexpectedly, however, it also over-penalizes those matched words in exactly the same word sequence as in the reference, i.e., a matched phrase. Therefore an adjustment is needed to extend the unit of position disorder penalty from word to phrase, which is defined as a sequence of consecutive matches between a candidate and a reference translation, so as to recognize the correct word order within a phrase. In practice, the phrase matching is performed as maximum matching in a way that each time the longest match is extracted until no more match can be found. In case of multiple phrases of the same size, the one with the shortest position distance is selected. This approach applies to the situation of multiple references as well, i.e., the longest phrase with the shortest position distance in any reference is selected first.

There is also a refinement of the penalty for word disorder, which is now further divided into *position distance* and *order distance*. The former is the original position distance of matched words, which is also extended to phrases now, to represent the distance of a matched phrase between its corresponding positions in the candidate and the reference. The latter concerns a sequence of matched phrases. It is quantified by assigning an index to each matched phrase sequentially, and then calculating the differences of the two indices for each match. The difference between these two types of distances is illustrated in Example 5. While both (1) and (2) share the same position distance, the matches give a correct sequence in (1) but a cross in (2), resulting in a bigger order distance for (2). It is expected that the different aspects of the variance of word order can be quantified more objectively and accurately this way.

Example 5



This was an assumption that word position could be better represented in terms of their relative positions in a sentence through the normalization of their position indices (see Example 3 above). Unfortunately, it was falsified by our subsequent experiments. Instead, using absolute positions lead to a higher correlation with human judgment. Therefore, the final adjustment of the distance penalty is, surprisingly, to shift relative index back to absolute index. Our observation is that the length-independent attribute of the relative index tends to suppress the effect of sentence length upon word order divergence. Example 6 below illustrates a comparison of the absolute and relative index with a short and a long sentence.[3] The words "international" in Candidate 1 and "countries" in Candidate 2 show similar position distances from their respective counterparts in the references in term of relative indices, i.e., 0.47 vs. 0.42, respectively. On the contrary, however the difference is several times greater if using absolute index, in that the two counterparts for the former are 3 words apart, whereas those for the latter are 14. Indeed, the word order divergence in Candidate 1 does not hinder our view in general, but that in Candidate 2 does pose a serious problem, due to the excessive length in this situation to magnify the interference effect of the word order divergence. Therefore, a better choice is to resort to the absolute index in regard to the sentence length in question.

Example 6 (Notation: [absolute|relative]):

Candidate 1: Short[1|.2] and[2|.4] various[3|.6] **international**[4|.8] news[5|1.0]
Reference: **International**[1|.33] news[2|.66] brief[3|1.0]
Candidate 2: Is[1|.03] on[2|.06] a[3|.09] popular[4|.13] the[5|.16] very[6|.19] in[7|.22] Iraq[8|.25] to[9|.28] those[10|.31] just[11|.34] like[12|.38] other[13|.41] world[14|.44] in[15|.47] which[16|.5] young[17|.53] people[18|.56] with[19|.59] the[20|.63] and[21|.66]flowers[22|.69] while[23|.72] awareness[24|.75] by[25|.78] other[26|.81] times[27|.84] of[28|.88] the[29|.91] **countries**[30|.94] of[31|.97] the[32|1.0]
Reference: Valentine's[1|.03] day[2|.06] is[3|.1] a[4|.13] very[5|.16] popular[6|.19] day[7|.23] in[8|.26] Iraq,[9|.29] as[10|.32] it[11|.35] is[12|.39] in[12|.42] the[14|.45] other[15|.48] **countries**[16|.52] of[17|.55] the[18|.58] world.[19|.61] Young[20|.65] men[21|.68] exchange[22|.71] with[23|.74] their[24|.77] girlfriends[25|.81] sweets,[26|.84] flowers,[27|.87] perfumes[28|.90] and[29|.94] other[30|.97] gifts.[31|1.0]

3.4 ATEC details

This section details the computation of the improved ATEC. Given a system candidate and a reference, their matched phrases are first retrieved. Each of them contains one or more matched words in a consecutive order, and has a phrase score computed as

$$\text{PhraseScore} = \sum \left(\text{Word}_{\text{type}} - \frac{w_{\text{info}}}{tfidf} \right) - \text{DisPenalty}$$

[3] Selected from NIST MetricsMATR08 development data (LDC2009T05).

where $\text{Word}_{\text{type}}$ refers to the score of a matched word, which depends on the type of match. This word score is further weighted by its information load, where w_{info} is a weight factor for the *tfidf* score of the matched word in question. For each phrase, there is a distance penalty DisPenalty comprising of position distance and order distance, to be computed as

$$\text{DisPenalty} = w_{\text{pos}} \times \text{PosDis} + w_{\text{ord}} \times \text{OrderDis}$$

where w_{pos} and w_{ord} are the weight factors for the position distance PosDis and the order distance OrderDis, respectively. There is an upper limit $\text{Limit}_{\text{dis}}$ for the distance penalty, which is a proportion of the sum of word scores within a phrase. When all phrase scores are available, an overall match score M can be calculated as

$$M = \sum \text{PhraseScore} - \text{InfoPenalty}_{\text{unmatch}}$$

where $\text{InfoPenalty}_{\text{unmatch}}$ refers to the information load of the unmatched reference words. It is approximated as

$$\text{InfoPenalty}_{\text{unmatch}} = \sum \left(\text{Word}_{\text{unmatch}} - \frac{w_{\text{info}}}{tfidf} \right)$$

where $\text{Word}_{\text{unmatch}}$ refers to the score for an unmatched word. There is an upper limit $\text{Limit}_{\text{info}}$ for the information penalty for the unmatched words, which is a proportion of $\text{InfoPenalty}_{\text{unmatch}}$.

The calculated match score M is then used to calculate the precision P and recall R, and their average F-measure for the following ATEC score for a sentence

$$\text{ATEC} = \frac{2PR}{P + R}$$

where P and R are defined in 3.2 above.

We further derived the optimized values for the parameters involved the ATEC calculation using the development data of NIST MetricsMATR08 with adequacy assessments by a simple hill climbing method. The optimal parameter setting is presented as in Table 1 below.

4 Evaluation

The performance of the improved ATEC metric is evaluated through two sets of experiments. The merits of different features incorporated into ATEC are first examined using two corpora from LDC, namely, Multiple-Translation Chinese (MTC) part 2 (LDC2003T17) and part 4 (LDC2006T04), which in total consist of 8,148 segments of Chinese-to-English MT outputs, with human assessments of adequacy and fluency. Table 2 presents the changes of correlation at segment level of adding each feature into ATEC, in Pearson correlation and 95% confidence interval (CI). Using surface

Table 1 Optimal values for ATEC parameters resulted from MetricsMATR development data

Parameters	Values
$\text{Word}_{\text{type}}$	1 (exact match), 0.95 (stem/sound/synonym match)
$\text{Word}_{\text{unmatch}}$	0.25
w_{info}	0.3
w_{pos}	0.03
w_{ord}	0.14
$\text{Limit}_{\text{dis}}$	0.95
$\text{Limit}_{\text{info}}$	0.4

Table 2 Merits of different features to ATEC on MTC corpora

	Multiple reference		Single reference	
	Ade (95% CI)	Flu (95% CI)	Ade (95% CI)	Flu (95% CI)
Matching modules				
Word	.372 (.353/.391)	.220 (.199/.240)	.333 (.314/.352)	.204 (.184/.225)
+ Stem	.395 (.377/.414)	.223 (.202/.244)	.361 (.342/.380)	.215 (.194/.236)
+ Sound	.393 (.374/.411)	.220 (.199/.240)	.361 (.342/.379)	.213 (.192/.234)
+ Sense	.396 (.378/.414)	.219 (.198/.238)	.373 (.354/.391)	.221 (.200/.241)
Information weight				
+ Info $_{\text{match}}$	.418 (.400/.436)	.234 (.213/.254)	.382 (.363/.400)	.226 (.206/.247)
+ Info $_{\text{unmatch}}$	.424 (.406/.442)	.233 (.212/.253)	.384 (.365/.402)	.228 (.202/.243)
Word order distance				
+ Position $_{\text{relative}}$	.432 (.414/.450)	.248 (.228/.269)	.377 (.358/.395)	.226 (.205/.246)
+ Position $_{\text{absolute}}$	**.445 (.428/.462)**	.264 (.244/.284)	**.393 (.374/.411)**	.242 (.222/.263)
+ Order (= ATEC $_{\text{improved}}$)	.438 (.421/.456)	**.293 (.273/.313)**	.385 (.366/.403)	**.258 (.237/.278)**
Baseline metrics				
ATEC original	.314 (.294/.333)	.171 (.150/.192)	.264 (.244/.284)	.136 (.114/.157)
BLEU-1	.382 (.363/.400)	.240 (.220/.261)	.346 (.327/.365)	.200 (.179/.220)
METEOR	.428 (.410/.446)	.277 (.257/.297)	.390 (.371/.408)	.237 (.217/.258)

Highest correlations of each group are marked as bold

word matching as a basis, the contribution of each feature is presented as the change of correlation it brings to the metric. Note that the inclusion of position distance in terms of absolute versus relative index is mutually exclusive. They are listed to illustrate the difference of the two ways of indexing. Three baseline metrics are provided for comparison, including the original ATEC metric. BLEU-1 and METEOR are known as precision- versus recall-oriented, respectively, in contrast to ATEC which is designed for a better balance by using F-measure.

The results in Table 2 show that most ATEC features do contribute positively to its performance, but in different manners. The multiple matching modules together with information weights lead to about 14% correlation improvement on adequacy in both

the multiple and single reference groups upon solely word matching (i.e., .372 → .424 and .333 → .384), two times as much as the 7% improvement on fluency on average. In contrast, the word order distance only brings a minor further improvement on adequacy, but a significant one on fluency (i.e., .233 → .293 (+26%) and .228 → .258 (+13%)). This confirms our view that in general the word choice and word order influences the adequacy and fluency respectively. In particular, the behaviors of some features are also worth noting. One is the merit of position distance with relative index, although positive, is significantly outweighed by the shift to absolute index. Another one is an interesting behavior of order distance: it reduces the correlation on adequacy but improves that on fluency.

Another set of experiments are carried out to further evaluate ATEC on MetricsMATR evaluation dataset. Table 3 presents the Pearson (segment level) and the Spearman (system level) correlations of the improved ATEC, together with its original version,[4] and three baseline metrics in different human assessment types.[5] It is confirmed that our further elaborations upon the original ATEC have substantially enhanced its performance, especially at the segment level, as demonstrated by the increase of correlation in all related groups of assessment. We can see that the improved ATEC has a performance favorably comparable to that of the few best metrics ranking at the top in MetricsMATR08. At the system level, however, the improved ATEC underperforms its original version occasionally in some groups, for instance, 4-point adequacy. Nevertheless, it is worth noting that the wide confidence intervals for these groups, e.g., .046/.858 in the group of 4-point adequacy for the improved ATEC as well as METEOR, may signify that the correlations at the system level may not be as reliable as they are at the segment level.

5 Conclusion

In this paper we have attempted to understand the key rationale for automatic MT evaluation from the perspective of the construction of meaning in text comprehension. Although this cognitive process is complex, the word choice and word order as two fundamental textual features are known for sure to play a key role in such construction of meaning for a sentence. Based on this observation we have proposed a novel evaluation metric ATEC[6] for MT evaluation with explicit measurement of these two features. It is formulated to capture the word choice by multiple matching modules and a quantification of word informativeness for both matched and unmatched words, and to assess the word order quantitatively by a penalty for the word position distance and the discordance of word sequence. While the relative contribution of each of these elements to the ATEC scoring is unknown, we opt for an empirical training to determine their optimal weights. Our experimental results confirm that the performance

[4] In the official evaluation results, there are four versions of ATEC in different parameter settings. The original ATEC here refers to ATEC2: with WordNet, vs. ATEC1: direct match only, ATEC3: with WordNet and preserve punctuation, and ATEC4: with WordNet and stoplist.

[5] For detail description of these assessment types, please refer to the official website of MetricsMATR at http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2008/results/correlationResults.html.

[6] The ATEC package can be downloaded from http://mega.ctl.cityu.edu.hk/ctbwong/ATEC/.

Table 3 Correlations of ATEC on MetricsMATR evaluation dataset (95% CI)

	ATEC $_{\text{improved}}$	ATEC $_{\text{original}}$	BLEU $_{\text{v11b}}$	METEOR $_{\text{v0.6}}$	TER $_{\text{v0.7.25}}$
7-point Ade					
Seg $_{\text{multiple}}$	.718 (.711/.726)	.649 (.640/.658)	.518 (.507/.529)	**.733 (.726/.740)**	−.535 (−.546/−.524)
Seg $_{\text{single}}$	.682 (.676/.689)	.580 (.572/.589)	.441 (.432/.451)	**.686 (.679/.692)**	−.524 (−.533/−.515)
Sys $_{\text{multiple}}$	.828 (.722/.897)	**.902 (.837/.942)**	.897 (.829/.939)	.846 (.748/.910)	−.895 (−.938/−.826)
Sys $_{\text{single}}$	.850 (.780/.899)	.838 (.763/.891)	.848 (.777/.898)	**.888 (.834/.925)**	**−.888(−.925/−.834)**
4-point Ade					
Seg $_{\text{multiple}}$	.711 (.679/.739)	.588 (.547/.626)	.459 (.409/.505)	**.721 (.690/.749)**	−.594 (−.632/−.553)
Seg $_{\text{single}}$	**.650 (.614/.684)**	.530 (.485/.572)	.350 (.296/.403)	.620 (.581/.656)	−.506 (−.550/−.459)
Sys $_{\text{multiple}}$	.582 (.046/.858)	.813 (.475/.942)	.802 (.450/.938)	.582 (.046/.860)	**−.835 (−.949/−.526)**
Sys $_{\text{single}}$	.676 (.199/.894)	.813 (.475/942)	.791 (.426/.935)	.588 (.055/.860)	**−.824 (−.946/−.500)**
Yes/No Ade					
Seg $_{\text{multiple}}$	.575 (.564/.585)	.498 (.486/.509)	.435 (.422/.447)	**.582 (.572/.592)**	−.432 (−.444/−.419)
Seg $_{\text{single}}$	**.551 (.543/.560)**	.463 (.453/.472)	.382 (.371/.392)	.548 (.539/.556)	−.423 (−.433/−.413)
Sys $_{\text{multiple}}$	.860 (.770/.916)	.791 (.666/.873)	**.866 (.781/.920)**	**.866 (.780/.920)**	−.792 (−.874/−.666)
Sys $_{\text{single}}$	.827 (.748/.883)	.765 (.662/.839)	.834 (.757/.888)	**.853 (.784/.901)**	−.816 (−.875/−.732)
Preference					
Seg $_{\text{multiple}}$	**.395 (.371/.418)**	.360 (.336/.383)	.249 (.223/.275)	.368 (.344/.392)	−.237 (−.263/−.211)
Seg $_{\text{single}}$	.336 (.317/.355)	.305 (.285/.324)	.251 (.231/.272)	**.337 (.318/.356)**	−.253 (−.273/−.233)
Sys $_{\text{multiple}}$	.729 (.575/.833)	.690 (.520/.808)	**.743 (.595/.843)**	.721 (.564/.828)	−.676 (−.798/−.500)
Sys $_{\text{single}}$	.677 (.546/.776)	.610 (.460/.726)	.681 (.551/.779)	**.691 (.564/.786)**	−.649 (−.755/−.509)
Concept					
Seg $_{\text{multiple}}$	**.712 (.681/.741)**	.574 (.531/.613)	.434 (.384/.482)	.706 (.674/.735)	−.520 (−.563/-.475)
Seg $_{\text{single}}$	**.631 (.593/.666)**	.507 (460/.550)	.333 (.278/.386)	.619 (.580/.655)	−.461 (−.508/−.412)

Table 3 continued

	ATEC $_{\text{improved}}$	ATEC $_{\text{original}}$	BLEU $_{\text{v11b}}$	METEOR $_{\text{v0.6}}$	TER $_{\text{v0.7.25}}$
Sys $_{\text{multiple}}$	.692 (.229/.900)	**.868 (.608/.960)**	.846 (.553/953)	.692 (.229/.900)	−.863(−.958/−.594)
Sys $_{\text{single}}$	.742 (.323/.918)	**.868 (.608/.960)**	.835 (.526/.949)	.698 (.239/.902)	−.835(−.949/−.526)
HTER					
Seg $_{\text{single}}$	.490 (−.512/−.467)	−.391 (−.416/−.366)	−.417 (−.441/−.393)	−.488 (−.510/−.466)	**.510 (.488/.531)**
Sys $_{\text{single}}$	−.708 (−.876/−.387)	−.675 (−.861/−.332)	−.753 (−.897/−.466)	**−.758 (−.899/−.475)**	.686 (.349/.866)

Highest correlations of each group are marked as bold

of ATEC is impressive, and favorably comparable to that of the other state-of-the-art evaluation metrics.

Nevertheless, there are still many aspects of the behavior of this metric remaining unknown requiring further studies, under different situations. Apparently it is unrealistic that the problem of MT evaluation could be satisfactorily resolved via a nice measurement of word choice and word order. The final remark in the above-mentioned early study of MT evaluation by Miller and Beebe-Center (1956), concerning these two features for assessing MT systems is that they are "useful to discriminate against very poor translations, but the present evidence indicates that it may not discriminate accurately in the range that might be labeled 'good' to 'excellent'." This observation seems to remain valid even for the evaluation metrics in use today. This situation certainly calls for an in-depth exploration of more textual and linguistic features in relation to translation quality and an insightful understanding of the mechanism how these features interplay to contribute to our text comprehension.

Acknowledgements The work described in this paper is supported by City University of Hong Kong through the Strategic Research Grant (SRG) 7002267. We would like to thank Mark Przybocki and NIST for helping us to run the ATEC package on their side for an authoritative evaluation on the MetricsMATR evaluation dataset. We also thank two anonymous reviewers for their insightful comments that help improve this paper a lot. The authors are nevertheless responsible for all remaining errors.

References

Babych B, Hartley A (2004) Extending the BLEU MT evaluation method with frequency weightings. In: Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-2004), Barcelona, Spain, 21–26 July 2004, pp 621–628

Banerjee S, Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, University of Michigan, Ann Arbor, MI, 29 June 2005, pp 65–72

Doddington G (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceeding of the second conference on human language technology (HLT-2002), San Diego, CA, 24–27 March 2002, pp 138–145

Giménez J, Màrquez L (2007) Linguistic features for automatic evaluation of heterogenous MT systems. In: Proceedings of the second workshop on statistical machine translation, Prague, Czech Republic, 23 June 2007, pp 256–264

Gopen GD (2004) The sense of structure: writing from the reader's perspective. Longman, New York

Krauwer S (1993) Evaluation of MT systems: a programmatic view. Mach Transl 8(1–2):59–66

Landauer TK (2002) On the computational basis of learning and cognition: arguments from LSA. In: Ross BH (ed) The psychology of learning and motivation, vol 41. Academic Press, New York, pp 43–84

Liu D, Gildea D (2005) Syntactic features for evaluation of machine translation. In: Proceedings of the workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, University of Michigan, Ann Arbor, MI, 29 June 2005, pp 25–32

Miller GA, Beebe-Center JG (1956) Some psychological methods for evaluating the quality of translations. Mech Transl 3(3):73–80

Owczarzak K, Van Genabith J, Way A (2007) Dependency-based automatic evaluation for machine translation. In: Proceedings of SSST, NAACL-HLT 2007/AMTA workshop on syntax and structure in statistical translation, Rochester, NY, 26 April 2007, pp 80–87

Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics (ACL-2002), Philadelphia, PA, pp 311–318

Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137

Russell RC (1918) US Patent 1,261,167, 2 April 1918
Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th conference of the association for machine translation in the Americas, visions of the future of machine translation (AMTA-2006), Cambridge, MA, USA, 8–12 August 2006, pp 223–231
White JS (2000) Contemplating automatic MT evaluation. In: White JS (ed) Proceedings of the 4th conference of the association for machine translation in the Americas, envisioning machine translation in the information future (AMTA-2000), Cuernavaca, Mexico, pp 100–108