A Simple Automatic MT Evaluation Metric

Petr Homola
Charles University
Prague, Czech Republic

Vladislav Kuboň
Charles University
Prague, Czech Republic

Pavel Pecina
Charles University
Prague, Czech Republic

{homola|vk|pecina}@ufal.mff.cuni.cz

Abstract

This paper describes a simple evaluation metric for MT which attempts to overcome the well-known deficits of the standard BLEU metric from a slightly different angle. It employes Levenshtein's edit distance for establishing alignment between the MT output and the reference translation in order to reflect the morphological properties of highly inflected languages. It also incorporates a very simple measure expressing the differences in the word order. The paper also includes evaluation on the data from the previous SMT workshop for several language pairs.

1 Introduction

The problem of finding a reliable machine translation metrics corresponding with a human judgment has recently returned to the centre of attention. After a brief period following the introduction of generally accepted and widely used metrics, BLEU (Papineni et al., 2002) and NIST (Doddington, 2002), when it seemed that this persistent problem has finally been solved, the researchers active in the field of machine translation (MT) started to express their worries that although these metrics are simple, fast and able to provide consistent results for a particular system during its development, they are not sufficiently reliable for the comparison of different systems or different language pairs.

The results of the NIST evaluation in 2005 (Le and Przybocki, 2005) have also strengthened the suspicion that the correlation between human judgment and the BLEU and NIST measures is not as strong as it was widely believed. Both measures seem to favor the MT output created by systems based on n-gram architecture, they are unable to take into account certain factors which are very important for the human judges of translation quality.

The article (Callison-Burch et al., 2006) thoroughly discusses the deficits of the BLEU and similar metrics. The authors claim that the existing automatic metrics, including some of the new and seemingly more reliable ones as e.g. Meteor (cf. (Banerjee and Lavie, 2005)) "... they are all quite rough measures of translation similarity, and have inexact models of allowable variation in translation." This claim is supported by a construction of translation variations which have identical BLEU score, but which are very different for a human judge. The authors identify three prominent factors which contribute to the inadequacy of BLEU – the failure to deal with synonyms and paraphrases, no penalties for missing content, and the crudeness of the brevity penalty.

Let us add some more factors based on our experiments with languages typologically different than English, Arabic or Chinese, which are probably the languages most frequently used in recent shared-task MT evaluations. The highly inflected languages and languages with a higher degree of word-order freedom may provide additional examples of sentences in which relatively small alterations of correct word forms may have a dire effect on the BLEU score while the sentence still remains understandable and acceptable for human evaluators.

The effect of rich inflection has been observed for example in (Týnovský, 2007), where the author mentions the fact that the BLEU score used for measuring the improvements in his experimental Czech-German EBMT system penalized heavily all subtle errors in Czech morphology arising from an out-of-context combined partial translations taken from different examples.

The problem of the insensitivity of BLEU to the variations of the order of n-grams identified in reference translations has already been mentioned in

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 33–36,
Athens, Greece, 30 March – 31 March 2009.

the paper (Callison-Burch et al., 2006). The authors showed examples where changing a good word order into an unacceptable one did not affect the BLEU score. We may add a different example documenting the phenomenon that a pair of syntactically correct Czech sentences with the same word forms, differing only in the word order whose n-gram score for $n = 2$, 3, and 4 differs greatly. Let us take one of the sentences from the 2008 SMT workshop and its reference translation:

When Caligula appointed his horse to the Senate, the horse at least did not have blood on its hoofs. — Když Caligula zvolil do senátu svého koně, neměl jeho kůň aspoň na kopytech krev.

If we modify the Czech reference sentence into *Když svého koně do senátu zvolil Caligula, jeho kůň aspoň neměl na kopytech krev.*, we destroy 8 out of 15 bigrams, 11 out of 14 trigrams and 12 out of 13 quadrigrams while we still have sentence with almost identical meaning and probably very similar human evaluation. The BLEU score of the modified sentence is, however, lower than it would be for the identical copy of the reference translation.

2 The description of the proposed metric

There is one aspect of the problem of a MT quality metric which tends to be overlooked but which is very important from the practical point of view. This aspect concerns the expected difficulties when post-editing the MT output. It is very important for everybody who really wants to use the MT output and who faces the decision whether it is better to post-edit the MT output or whether a new translation made by human translators would be faster and more efficient way towards the desired quality. It is no wonder that such a metric is mentioned only in connection with systems which really aim at practical exploitation, not with a majority of experimental MT system which will hardly ever reach the stage of industrial exploitation.

We have described one example of such practically oriented metric in (Hajič et al., 2003). The metric exploits the matching algorithm of Trados Translator's Workbench for obtaining the percentage of differences between the MT output and the reference translation (created by post-editing the MT output). The advantage of this measure is its close connection to the real world of human translating by means of translation memory, the disadvantage concerns the use of a proprietary matching algorithm which has not been made public and which requires the actual use of the Trados software.

Nevertheless, the matching algorithm of Trados gives results which to a great extent correspond to a much simpler traditional metric, to the Levenshtein's edit distance. The use of this metric may help to refine a very strict treatment of word-form differences by BLEU. A similar approach at the level of unigram matching has been used by the well-known METEOR metric (Agarwal and Lavie, 2008), which proved its qualities during the previous MT evaluation task in 2008 (Callison-Burch et al., 2008). Meteor uses Porter stemmer as one step in the word alignment algorithm. It also relies on synonymy relations in WordNet.

When designing our metric, we have decided to follow two general strategies – to use as simple means as possible and to avoid using any language dependent tools or resources. Levenshtein metric (or its modification for word-level edit distance) therefore seemed to be the best candidate for several aspects of the proposed measure.

The first aspect we have decided to include was the inflection. The edit distance has one advantage over the language independent stemmer – it can uniformly handle the differences regardless of their position in the string. The stemmer will probably face certain problems with changes inside the stem as e.g. in the Czech equivalent of the word *house* in different cases *dům* (nom.sg) — *domu* (gen., dat. or loc. sg.) or German *Mann* in different numbers *der Mann* (sg.) — *die Männer* (pl.), while the edit distance will treat them uniformly with the variation of prefixes, suffixes and infixes.

As mentioned above, we have also intended to aim at the treatment of the free word order in our metric. However this seems to be one of the major flaws of the BLEU score, it turned out that the word order is extremely difficult if we stick to the use of simple and language independent means. If we take Czech as an example of a language with relatively high degree of word-order freedom, we can still find certain restrictions (e.g. the sentence-second position of clitics, their mutual order, the adjectives typically, but not always preceding the nouns they depend upon etc.) which will definitely influence the human judgment of the acceptability of a particular sentence. These restrictions are language dependent (for example Polish, the

language very closely related to Czech, has different rules for congruent attributes, the adjectives stand much more often to the right of the governing noun) and they are also very difficult to capture algorithmically. If the MT output is compared to a single reference translation only, there is, in fact, no way how the metric could account for the possible correct variations of the word order without exploiting very deep language dependent information. If there are more reference translations, it is possible that they will provide the natural variations of the word order, but it, in fact, means that if we want to stick to the above mentioned requirements, we have to give up the hope that our metric will capture this important phenomenon.

2.1 Word alignment algorithm

In order to capture the word form variations caused by the inflection, we have decided to employ the following alignment algorithm at the level of individual word forms. Let us use the following notation: Let the reference translation **R** be a sequence of words r_i, where $i \in < 1, \ldots, n >$. Let the MT output **T** be a sequence of words tj, where $j \in < 1, \ldots, m >$. Let us also set a threshold of similarity $s \in < 0, 1 >$. (s roughly expresses how different the forms of a lemma may be. The idea behind this criterion is that a mistake in one morphological category (reflected mostly by a different ending of the corresponding word form) is not as serious as a completely different lexeme. This holds especially for morphologically rich languages that can have tens or even hundreds of distinct word forms for a single lemma.) Starting from t_1, let us find for each t_j the best r_i for $i \in < 1, \ldots, n >$ such that the edit distance d_j from t_j to r_i normalized by the length of t_j is minimal and at the same time $d_j < s$. If the r_i is already aligned to some t_k, $k < j$ and the edit distance $d_k > d_j$, then align t_j to r_i and re-calculate the alignment for t_k to its second best candidate, otherwise take the second best candidate r_l conforming with the above mentioned conditions and align it to t_j. As a result of this process, we get the alignment score A_{TR} from **T** to **R**. $A_{TR} = \frac{\sum(1-d_i)}{m}$ (for $i \in < 1, \ldots, n >$) where $d_i = 1$ for those word forms t_i which are not aligned to any of the word forms r_j from **R**. Then we calculate the alignment score A_{RT} using the same algorithm and aligning the words from **R** to **T**. The similarity score **S** equals the minimum from A_{TR} and A_{RT}. The way how the similarity score **S** is constructed ensures that the score takes into account a difference in length between **T** and **R**, therefore it is not necessary to include any brevity penalty into the metric.

2.2 A structural metric

In order to express word-order difference between the MT output and the reference translation we have designed a structural part of the metric. It is based on an algorithm similar to one of the standard sorting methods, an insert sort. The reference translation **R** represents the desired word order and the algorithm counts the number of operations necessary for obtaining the correct word order from the word order of the MT output **T** by inserting the words t_i to their desired positions r_j (t_i is aligned to r_j). If a particular word t_i is not aligned to any r_j, a penalty of 1 is added to the number of operations.

2.3 A combination of both metrics

The overall score is computed as a weighted average of both metrics mentioned above. Let L be the lexical similarity score and M the structural score based on a word mapping. Then then overall score S can be obtained as follows:

$$S = aL + bM$$

The coefficients a and b must sum up to one. They allow to capture the difference in the degree of word-order freedom among target languages. The coefficient b should be set lower for the target languages with more free word-order. Because both then partial measures L and M have values in the interval $< 0, 1 >$, the value of S will also fall into this interval.

3 The experiment

We have performed a test of the proposed metric using the data from the last year's SMT workshop.[1] The parameters a, b, and s have been set to the same value for all evaluated language pairs, no language dependent alterations were tested in this experiment:

Parameter	Value
s	0.15
a	0.9
b	0.1

[1] The data are available at http://www.statmt.org/wmt08.

The values for the parameters have been set up empirically with special attention being paid to Czech, the only language with really rich inflection among the languages being tested.

We have performed sentence-level and system-level evaluation using the Spearman's rank correlation coefficient which is defined as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where $d_i = x_i - y_i$ is the difference between the ranks of corresponding values X_i and Y_i and n is the number of values in each data set.

The following scores express the correlation of our automatic metric and the human judgements for the language pairs English-Czech and English-German. The sentence-level correlation ρ_{sent} is the average of Spearman's ρ across all sentences.

Language pair	Metric	ρ_{sent}	ρ_{sys}
English-Czech	proposed	0.20	0.50
English-Czech	BLEU	0.21	0.50
English-German	proposed	0.91	0.37
English-German	BLEU	0.90	0.20

3.1 Conclusions

The metric presented in this paper attempts to combine some of the important factors which seem to be neglected by some generally accepted MT evaluation metrics. Inspired by the fact that human judges tend to accept incorrect word-forms of corectly translated lemmas, it employs a similarity measure relaxing the requirements on identity (or similarity) of matching word forms in the MT output and the reference translation. At the same time, it also incorporates a penalty for different length of the MT output and the reference translation. The second component of the metric tackles the problem of incorrect word-order. The constants used in the metric allow to set the weight of its two components with regard to the target language properties.

The experiments performed on the data from the previous shared evaluation task are promising. They indicate that the first component of the metric succesfully replaces the strict unigram measure used in BLEU while the second component may require certain alteration in order to achieve a higher correlation with human judgement.

Acknowledgments

The presented research has been supported by the grant No. 1ET100300517 of the GAAV ČR and by Ministry of Education of the Czech Republic, project MSM 0021620838.

References

Abhaya Agarwal and Alon Lavie. 2008. *Meteor, M-BLEU and M-TER: Evaluation metrics for high correlation with human rankings of machine translation output*. In Proceedings of the Third Workshop on Statistical Machine Translation, pages 115-118. Columbus, Ohio, Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. *Meteor: An automatic metric for MT evaluation with improved correlation with human judgments.*. In Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, Ann Arbor, Michigan.

Chris Callison-Burch, Miles Osborne, Philipp Koehn. 2006. *Re-evaluating the Role of BLEU in Machine Translation Research.*. In Proceedings of the EACL'06, Trento, Italy.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, Josh Schroeder. 2008. *Further Meta-Evaluation of Machine Translation.*. In Proceedings of the Third Workshop on Statistical Machine Translation, pages 70-106, Columbus, Ohio. Association for Computational Linguistics.

George Doddington. 2002. *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.* In Proceedings of the second international conference on Human Language Technology Research, San Diego, California,USA

Jan Hajič, Petr Homola, Vladislav Kuboň. 2003. *A Simple Multilingual Machine Translation System.*. In Proceedings of the MT Summit IX, New Orleans, USA.

Kishore Papineni, Salim Roukos, ToddWard, and Wei-Jing Zhu. 2002. *Bleu: A method for automatic evaluation of machine translation.*. In Proceedings of ACL 2002.

Audrey Le and Mark Przybocki. 2005. *NIST 2005 machine translation evaluation official results.*. Official release of automatic evaluation scores for all submissions.

Miroslav Týnovský. 2007. *Exploitation of Linguistic Information in EBMT.*. Master thesis at Charles University in Prague, Faculty of Mathematics and Physics.