The Back-translation Score: Automatic MT Evaluation at the Sentence Level without Reference Translations

Reinhard Rapp
Universitat Rovira i Virgili
Avinguda Catalunya, 35
43002 Tarragona, Spain
reinhard.rapp@urv.cat

Abstract

Automatic tools for machine translation (MT) evaluation such as BLEU are well established, but have the drawbacks that they do not perform well at the sentence level and that they presuppose manually translated reference texts. Assuming that the MT system to be evaluated can deal with both directions of a language pair, in this research we suggest to conduct automatic MT evaluation by determining the orthographic similarity between a back-translation and the original source text. This way we eliminate the need for human translated reference texts. By correlating BLEU and back-translation scores with human judgments, it could be shown that the back-translation score gives an improved performance at the sentence level.

1 Introduction

The manual evaluation of the results of machine translation systems requires considerable time and effort. For this reason fast and inexpensive automatic methods were developed. They are based on the comparison of a machine translation with a reference translation produced by humans. The comparison is done by determining the number of matching word sequences between both translations. It could be shown that such methods, of which BLEU (Papineni et al., 2002) is the most common, can deliver evaluation results that show a high agreement with human judgments (Papineni et al., 2002; Coughlin, 2003; Koehn & Monz, 2006).

Disadvantages of BLEU and related methods are that a human reference translation is required, and that the results are reliable only at corpus level, i.e. when computed over many sentence pairs (see e.g. Callison-Burch et al., 2006). However, at the sentence level, due to data sparseness the results tend to be unsatisfactory (Agarwal & Lavie, 2008; Callison-Burch et al., 2008). Papineni et al. (2002) describe this as follows: "BLEU's strength is that it correlates highly with human judgments by averaging out individual sentence judgment errors over a test corpus rather than attempting to divine the exact human judgment for every sentence: *quantity leads to quality.*"

Although in many scenarios the above mentioned drawbacks may not be a major problem, it is nevertheless desirable to overcome them. This is what we attempt in this paper by introducing the *back-translation score*. It is based on the assumption that the MT system considered can translate a language pair in both directions, which is usually the case. Evaluating the quality of a machine translation now involves translating it back to the source language. The score is then computed by comparing the back-translation to the original source text. Although for this comparison BLEU could be used, our experiments show that a modified version which we call *OrthoBLEU* is better suited for this purpose as it can deal with compounds and inflexional variants in a more appropriate way. Its operation is based on finding matches of character- rather than word-sequences. It resembles algorithms used in translation memory search for locating orthographically similar sentences.

The results that we obtain in this work refute to some extend the common belief that back-translation (sometimes also called round-trip translation) is not a suitable means for MT evaluation (Somers, 2005; Koehn, 2005). This belief seems to be largely based on the obvious observation that the back-translation score is highest for a trivial translation system that does nothing and simply leaves all source words in place. On the other hand, according to Somers (2005) "until now no one as far as we know has published results demonstrating this" (i.e. that back-translation is not useful for MT evaluation).

We would like to add that so far the inappropriateness of back-translation has only been shown by comparisons with other automatic metrics (Somers 2005; Koehn, 2005), which are also

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 133–136,
Suntec, Singapore, 4 August 2009. ©2009 ACL and AFNLP

flawed. Somers (2005) therefore states: “To be really sure of our results, we should like to replicate the experiments evaluating the translations using a more old-fashioned method involving human ratings of intelligibility.” That is, apparently nobody has ever seriously compared back-translation scores to human judgments, so the belief about their inutility seems not sufficiently backed by facts. This is a serious deficit which we try to overcome in this work.

2 Procedure

As our test corpus we use the first 100 English and German sentences of the *News Corpus* which was kindly provided by the organizers of the *Third Workshop on Statistical Machine Translation* (Callison-Burch et al., 2008). This corpus comprises human translations of articles from various news websites. In the case of the 100 sentences used here, the source language was Hungarian and the translations to English and German were produced from the Hungarian original. As MT evaluation is often based on multilingual corpora, the use of indirect translations appears to be a realistic scenario.

The 100 English sentences were translated to German using the online MT-system Babel Fish (`http://de.babelfish.yahoo.com/`) which is based on Systran technology. Subsequently, the translations were back-translated to English. Table 1 shows a sample sentence and its translations.

English (source)	The skyward zoom in food prices is the dominant force behind the speed up in eurozone inflation.
German (human translation)	Hauptgrund für den in der Eurozone gemessenen Anstieg der Inflation seien die rasant steigenden Lebensmittelpreise.
German (Babel Fish)	Die gen Himmel Lebensmittelpreise laut summen innen ist die dominierende Kraft hinter beschleunigen in der Eurozoneinflation.
English (back-translation)	Towards skies the food prices loud hum inside are dominating Kraft behind accelerate in the euro zone inflation.

Table 1: Sample sentence, its human translation, and its Babel Fish forward and backward translations.

The Babel Fish translations to German were judged by the author according to the standard criteria of *fluency* and *adequacy*. Hereby the scale provided by Koehn & Monz (2006) was used which assigns values between 1 and 5. We then for each sentence computed the mean of its fluency and adequacy values. This somewhat arbitrary measure serves the purposes of designating each sentence a single value, which makes the subsequent comparisons with automatic evaluations easier.

Having completed the human judgments, we next computed automatic judgments using the standard BLEU score. For this purpose we used the latest version (v12) of the NIST tool, which can be freely downloaded from the website `http://www.nist.gov/speech/tests/mt/`. This tool not only computes the BLEU score, but also a slightly modified variant, the so-called NIST score. Whereas the BLEU score assigns equal weights to all word sequences, the NIST score tries to take a sequence’s information content into account by giving less frequent word sequences higher weights. In addition, the so-called *brevity penalty*, which tries to penalize too short translations, is computed somewhat differently, with the effect that small length differences have less impact on the overall score.

Using the NIST tool, the BLEU and NIST scores for all 100 translated sentences where computed. Hereby, the human translations were taken as reference. In addition, the BLEU and NIST scores were also computed for the back-translations, thereby using the source sentences as reference.

By doing so we must emphasize that, as described in the previous section, the BLEU score was not designed to deliver satisfactory results at the sentence level (Papineni et al., 2002), and this also applies to the closely related NIST score. On the other hand, there are no simple automatic evaluation tools that are suitable at the sentence level. Only the METEOR-System (Agarwal & Lavie, 2008) is a step in this direction. It takes into account inflexional variants and synonyms. However, it is considerably more sophisticated and is highly dependent on the underlying large scale linguistic resources.

We also think that – irrespectively of their design goals – the performance of the established BLEU and NIST scores at the sentence level is of some interest, especially as to our knowledge no other quantitative figures have been published so far. For the current work, as improved evaluation at the sentence level is one of the goals, this appears to be the only possibility to at all provide some baseline for a comparison using a well established automatic system.

In an attempt to reduce the concerns that arise from applying BLEU at the sentence level, we introduce OrthoBLEU. Like BLEU OrthoBLEU also compares a machine translation to a reference translation. However, instead of word sequences sequences of characters are considered, as proposed by Denoual & Lepage (2005). The OrthoBLEU score between two strings is computed as the (relative) number of their matching triplets of characters (trigrams). Figure 1 illustrates this using the words *pineapple* and *apple pie*. As 6 out of 11 trigrams match, the resulting OrthoBLEU score is 54.5%.

The procedure illustrated in Figure 1 is not only applicable to words, but likewise to sentences, as punctuation marks, blanks, and special symbols can be treated like any other character. It is obvious that this procedure, which was originally developed for the purpose of fuzzy information retrieval, shows some tolerance with regard to inflexional variants, compounding, and derivations, which should be advantageous in the current setting. The source code of OrthoBLEU was written in C and can be freely downloaded from the following URL: `http://www.fask.uni-mainz.de/user/rapp/comtrans/`.

Using the OrthoBLEU algorithm, the evaluations previously conducted with the NIST tool were repeated. That is, both the Babel Fish translations as well as their back-translations were evaluated, whereby in the first case the human translations and in the second case the source sentences served as references.

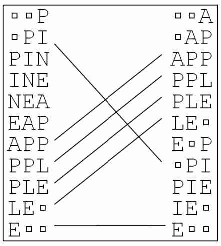

Figure 1: Computation of the OrthoBLEU score.

3 Results

Table 2 gives the average results of the evaluations described in the previous section. In columns 1 and 2 we find the human evaluation scores for fluency and adequacy, and column 3 combines them to a single score by computing their arithmetic mean. Columns 4 and 5 show the NIST and BLEU scores as computed using the NIST tool. They are based on the Babel Fish translations from English to German, whereby the human translations served as the reference. Column 6 shows the corresponding score based on OrthoBLEU, which delivers values in a range between 0% and 100%. Columns 7 to 9 show analogous scores for the back-translations. In this case the English source sentences served as the reference. As can be seen from the table, the values are higher for the back-translations. However, it would be premature to interpret this observation such that the back-translations are better suited for evaluation purposes. As these are very different tasks with different statistical properties, it would be methodologically incorrect to simply compare the absolute values. Instead we need to compute correlations between automatic and human scores.

This we did by correlating all NIST-, BLEU-, and OrthoBLEU scores for all 100 sentences with the corresponding (mean fluency/adequacy) scores from the human evaluation. We computed the Pearson product-moment correlation coefficient for all pairs, with the results being shown in Table 3. Hereby a coefficient of +1 indicates a direct linear relation, a coefficient of -1 indicates an inverse linear relation, and a coefficient of 0 indicates no linear relation.

When looking at the "translation" section of Table 3, as to be expected we obtain very low correlation coefficients for the BLEU and the NIST scores. This confirms their unsuitability for application at the sentence level as expected (see section 1). For the OrthoBLEU score we also get a very low correlation coefficient of 0.075, which means that OrthoBLEU is also unsuitable for evaluation of direct translations at the sentence level.

However, when we look at the back-translation section of Table 3, the situation is somewhat different. The correlation coefficient for the NIST score is still slightly negative, indicating that trying to take a word sequence's information content into account is hopeless at the sentence level. However, the correlation coefficient for the BLEU score almost doubles from 0.078 to 0.133, which, however, is still unsatisfactory. But a surprise comes with the OrthoBLEU score: It more than quadruples from 0.075 to 0.327, which at the sentence level is a rather good value as this result comes close to the correlation coefficient of 0.403 reported by Agarwal & Lavie (2008) as the very best of several values obtained for the METEOR system. Remember that, as described in section 2, the METEOR system requires a human-generated ref-

HUMAN EVALUATION			AUTOMATIC EVALUATION OF FORWARD-TRANSLATION			AUTOMATIC EVALUATION OF BACK-TRANSLATION		
FLUENCY	ADEQUACY	MEAN	NIST	BLEU	ORTHO-BLEU	NIST	BLEU	ORTHO-BLEU
2,49	3,06	2,78	1,31	0,01	39,72%	2,90	0,25	68,94%

Table 2: Average BLEU, NIST and OrthoBLEU scores for the 100 test sentences.

Translation	Human evaluation – NIST	-0,169
	Human evaluation – BLEU	0,078
	Human evaluation – OrthoBLEU	0,075
Back-translation	Human evaluation – NIST	-0,102
	Human evaluation – BLEU	0,133
	Human evaluation – OrthoBLEU	0,327

Table 3: Correlation coefficients between human and various automatic judgments based on 100 test sentences.

erence translation, large linguistic resources and comparatively sophisticated processing, and that all of this is unnecessary for the back-translation score.

4 Discussion and prospects

The motivation for this paper resulted from observing a contradiction: On one hand, practitioners sometimes recommend that (if one does not understand the target language) a back-translation can give some idea of the translation quality. Our impression has always been that this is obviously true for standard commercial systems. On the other hand, serious scientific publications (Somers, 2005; Koehn, 2005) come to the conclusion that back-translation is completely unsuitable for MT evaluation.

The outcome of the current work is in favor of the first point of view, but we should emphasize that we have no doubt about the correctness of the results presented in the publications. The discrepancy is likely to result from the following:

- The previous publications did not compare back-translation scores to human judgments but to BLEU scores only.
- The introduction of OrthoBLEU improved back-translation scores significantly.

What remains is the fact that evaluation based on back-translations can be easily fooled, e.g. by a system that does nothing, or that is capable of reversing errors. These obvious deficits have probably motivated reservations against such systems, and we agree that for such reasons they may be unsuitable for use at MT competitions.[1] However, there are numerous other applications where such considerations are of less importance. Also, it might be possible to introduce a penalty for trivial forms of translation, e.g. by counting the number of word sequences (e.g. of length 1 to 4) in a translation that are not found in a corpus of the target language.[2]

Acknowledgments

This research was in part supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme. We would also like to thank the anonymous reviewers for their comments, the providers of the NIST MT evaluation tool, and the organizers of the *Third Workshop on Statistical MT* for making available the *News Corpus*.

References

Abhaya Agarwal, Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. *Proc. of the 3rd Workshop on Statistical MT*, Columbus, Ohio, 115–118.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Josh Schroeder. 2008. Further meta-evaluation of machine translation. *Proc. of the 3rd Workshop on Statistical MT*, Columbus, 70–106.

Chris Callison-Burch, Miles Osborne, Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. *Proc. of 11th EACL*, 249–256.

Deborah Coughlin. 2003. Correlating automated and human assessments of machine translation quality. *Proc. of MT Summit IX, New Orleans*, 23–27.

Etienne Denoual, Yves Lepage. 2005. BLEU in characters: towards automatic MT evaluation in languages without word delimiters. *Proc. of 2nd IJCNLP, Companion Volume*, 81–86.

Philipp Koehn. 2005. Europarl: A parallel corpus for evaluation of machine translation. *Proceedings of the 10th MT Summit*, Phuket, Thailand, 79–86.

Philipp Koehn, Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. *Proc. of the Workshop on Statistical MT*, New York, 102–121.

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. Proc. of the *40th Annual Meeting of the ACL*, 311–318.

Harold Somers. 2005. Round-trip translation: what is it good for? In *Proceedings of the Australasian Language Technology Workshop ALTW 2005*. Sydney, Australia. 127–133.

[1] Although there might be a solution to this: It may not always be necessary that forward and backward translations are generated by the same MT system. For example, in an MT competition back-translations could be generated by all competing systems, and the resulting scores could be averaged.

[2] Looking up single words would not be sufficient as a system establishing any unambiguous 1:1 relationship between the source and the target language vocabulary would obtain top scores.