Calibrating resource-light automatic MT evaluation: a cheap approach to ranking MT systems by the usability of their output

Bogdan Babych
Centre for Translation Studies
University of Leeds
Leeds LS2 9JT, UK
bogdan@comp.leeds.ac.uk

Debbie Elliott
School of Computing
University of Leeds
Leeds LS2 9JT, UK
debe@comp.leeds.ac.uk

Anthony Hartley
Centre for Translation Studies
University of Leeds
Leeds LS2 9JT, UK
a.hartley@leeds.ac.uk

Abstract

MT systems are traditionally evaluated with different criteria, such as adequacy and fluency. Automatic evaluation scores are designed to match these quality parameters. In this paper we introduce a novel parameter – usability (or utility) of output, which was found to integrate both fluency and adequacy. We confronted two automated metrics, BLEU and LTV, with new data for which human evaluation scores were also produced; we then measured the agreement between the automated and human evaluation scores. The resources produced in the experiment are available on the authors' website.

Introduction

In recent years, considerable effort has been invested in developing and testing methods of automatically evaluating the quality of MT output that correlate reliably with human judgments. However, such methods tend to be resource-heavy insofar as they require significant amounts of training or reference data.

This paper describes a comparative evaluation of two mature knowledge-based MT systems, based on human judgments of three quality attributes, designed to calibrate two resource-light automatic methods.

Since the impetus given to research into automatic evaluation of MT output quality by (Brew and Thompson, 1994), it is the BLEU[1] approach (Papineni et al., 2002) that has enjoyed the widest uptake. The RED method (Akiba et al., 2003) ranks texts and takes an edit-distance approach over PoS-tagged data which, it is claimed, handles long-distance co-occurrence and is less sensitive than BLEU to the choice of reference translations. However, the test suite and training data are again acknowledged as being expensive to produce. (Rajman and Hartley, 2001; 2002) propose a method combining syntactic relations and semantic vectors that dispenses with the need for reference translations but which requires parsed data and a large aligned training corpus.

We have applied both the BLEU and the LTV metric (Babych, 2004) to a corpus of business texts translated from French into English by two mature knowledge-based MT systems, with a view to scoring the systems. We also sought to establish whether the quality of the translations was judged by humans to be improved by updating the dictionaries of each system in line with a benchmark provided by a human translation, and whether the automated metrics would capture this perception.

Automatic evaluation – BLEU method

The BLEU automatic evaluation metric has been shown to strongly correlate with human judgements about fluency of knowledge-based MT systems, which is also confirmed by the results presented here. The BLEU method is based on matches of N-grams (individual words or sequences of several words, usually up to 4) in MT and in one or more human "gold standard" reference translations. More specifically, BLEU measures N-gram precision (the proportion of N-grams found both in MT output and in any of the "gold standard" human reference translations).

The rationale of using BLEU is to explore objective properties of the evaluated texts as compared to a gold standard human reference translation. This gives an "absolute" measure for comparison across different evaluation attributes, e.g. fluency, adequacy and usability, which are not directly comparable through human scoring. The BLEU scores are in the range [0, 1].

Automatic evaluation – LTV method

The LTV (Legitimate Translation Variation) method as described in (Babych, 2004) is based on BLEU, but the matched words in the tested MT output and the "gold standard" translation have unequal weight when they are matched. More weight is given to statistically significant words in the evaluated text. Statistical significance weights, suggested in (Babych, Hartley, Atwell, 2003) are computed by contrasting the word's frequency in a text and in the rest of the corpus: the formula is similar to the tf.idf score used in Information Retrieval, but the scores are normalised by the relative frequency of the word in the corpus.

Usually the content words, names of events, event participants, and terminology happen to be more statistically significant. The intuition is that such words normally have a unique translation equivalent, whereas functional words and other words, which less frequent in a given text than in the rest of the corpus, are subject to greater Legitimate Translation Variation, i.e. they will vary across independently produced human translations of the same text. Therefore, matches of the "significant" words should count more, when the MT output is evaluated, which is captured in LTV method by assigning greater weights to words whose statistical significance score is >1.

LTV computes three scores for each evaluated document: Precision (or degree of avoiding "over-generation" of "significant" words), recall (or degree of avoiding "under-generation") and F-score, where precision and recall are weighted equally. In our previous experiments with the DARPA corpus, *recall* was found to be the best match for *adequacy*, and the *F-score* for *fluency*.

[1] BLEU stands for BiLingual Evaluation Understudy

Calibrating BLEU and LTV

Set up of the experiment

We evaluated the French-to-English versions of two leading commercial MT systems – System 1 and System 2 – in order to assess the quality of their output and to determine whether updating the system dictionaries brought about an improvement in performance.

The input for the evaluation were a White Paper (3,334 words in 120 segments) from the European Commission and a collection of 36 business and personal emails (average length 107 words). We also had translations of all the texts by a professional translator. We used these as a gold standard reference for creating new dictionary entries. These human translations also figured in the evaluation exercise.

For the emails, we also had translations produced by a non-professional, French-speaking translator. This was intended to simulate a situation where, in the absence of MT, the author of the email would have to write in a foreign language (here English). We anticipated that the quality would be judged lower than the professional, native speaker translations.

The evaluations were performed by 70 judges – 42 business people and 28 postgraduate students who knew very little or no French.

Using a five-point scale in each case, judgments were solicited on three attributes of text quality by means of the following questions:

- usability – "Using each reference email on the left, rate the three alternative versions on the right according to **how usable you consider them to be for getting business done**." The non-native translations were dispersed anonymously in the data set and so were also judged.
- fluency – "Look carefully at each segment of text and give each one a score according to how much you think the text reads like fluent English written by a native speaker." No reference text was seen.
- adequacy – "For each segment, read carefully the reference text on the left. Then judge how much of the same content you can find in the candidate text."

Five independent judgments were collected for each segment and for each email.

Human evaluation results

Figure 1 and Table 1 summarise the results of human evaluation for 3 different evaluation tasks:

1. fluency of the White Paper translations (the 2 MT systems before and after dictionary update), judged by students (40%) and business users (60%) – FLU.

2. adequacy of the White Paper translation (the 2 MT systems before and after dictionary update), judged by students – ADE.

3. usability of the email translations (the 2 MT systems before and after dictionary update and a non-native speaker translation), judged by business users – USL.

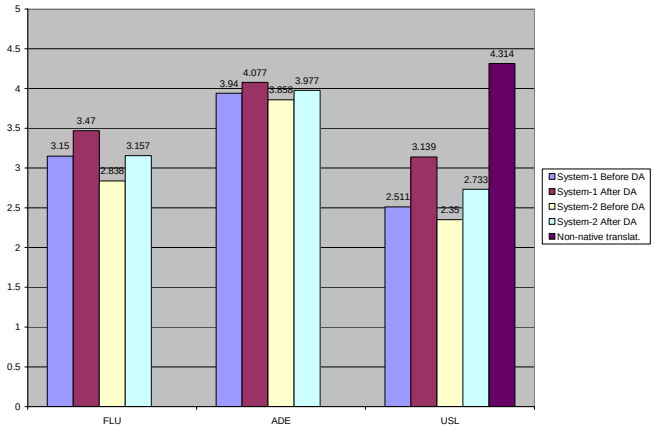

Figure 1. Human evaluation results

	FLU	ADE	USL
System-1 Before DA	3.15	3.94	2.511
System-1 After DA	3.47	4.077	3.139
System-2 Before DA	2.838	3.858	2.35
System-2 After DA	3.157	3.977	2.733
Non-native translation			4.314

Table 1. Human evaluation results

It can be seen from the figures that the results for adequacy are very high: on average MT systems scored "four" on the five-point scale. The results for fluency are worse: "three" on the five-point scale is the most likely score for MT systems. This shows that MT is useful primarily for "assimilation", i.e., "understanding" purposes, where the users try to grasp the meaning, and are less interested in getting well-formed, i.e., grammatically and lexically impeccable and stylistically natural sentences (which might be important for "dissemination", e.g., publication purposes – for these tasks MT is still not so good).

On the other hand, usability most probably has integrated "fluency" and "adequacy" aspects of the text quality (and perhaps has been influenced by the presence of the non-native human translation). It is natural to suggest that the text which is easier to read requires less effort on the part of the user to reconstruct the meaning. From the point of view of usability, fluency and adequacy MT errors aggravate each other, so the scores for usability are lower than for the other two attributes.

All human scores for texts *after dictionary update* are consistently higher both for System 1 and for System 2, but the degree of improvement is different: it is the biggest for usability of the e-mail translations (25% for System 1 and 16% for System 2), and the smallest for adequacy of the whitepaper translation (3.5% for System 1 and 3.1% for System 2).

Automatic evaluation results

The results of BLEU evaluation for the whitepaper document and for emails are summarised in Figure 2. BLEU used a single human reference translation and counted N-grams up to N=4.

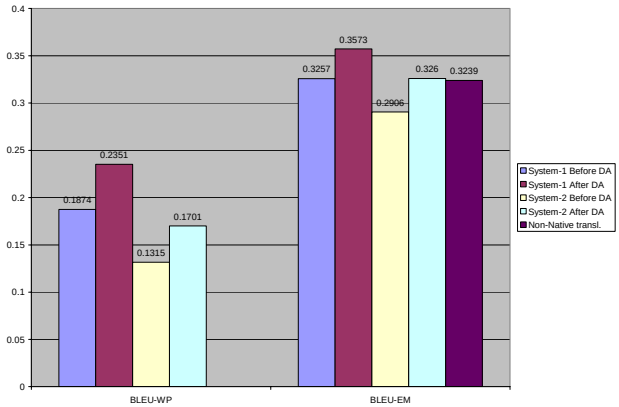

Figure 2. BLEU evaluation: White Paper and emails

	BLEU-WP	BLEU-EM
System-1 Before DA	0.1874	0.3257
System-1 After DA	0.2351	0.3573
System-2 Before DA	0.1315	0.2906
System-2 After DA	0.1701	0.3260
Non-Native transl.		0.3239

Table 2. BLEU evaluation: White Paper and emails

Another aspect of the BLEU evaluation is a possible comparison between the White Paper text and in the business emails. There are many more matches of N-grams in the emails as compared to the White Paper. Table 3 summarises the growth of matches between these two types of documents.

System-1 Before DA	0.737994
System-1 After DA	0.519779
System 2 Before DA	1.209886
System 2 After DA	0.916520

Table 3. Percentage growth of N-gram matches in the emails over the White Paper

The table shows that translating emails is objectively easier for MT systems than translating the legal documents. However, human judges adjust the scores according to the evaluation task, so the difference becomes apparent only with automatic evaluation. In our experiment, since the human non-native translation was involved in usability evaluation of the emails, a kind of "masking effect" was introduced, so the scores for usability were lower than for adequacy or fluency (where there was no comparison with the human translation). Therefore the BLEU score allows us to make comparison between different types of texts, which were not directly compared in our evaluation and shows that translating emails is easier for MT systems and much better results are objectively achievable, in comparison to the legal documents.

Also in Table 3 the difference between the whitepaper and the email matches for System 1 is lower than for System 2 (74% and 52% vs. 121% and 91%). This shows that System 1 translation gives more stable quality across genres, and the quality for System 2 is more dependent on the genre of the translated text: it achieves its quality is greatly improved for "easier" texts, such as emails as compared to the "hard" texts.

Figure 3 and Table 4 summarise the LTV evaluation results.

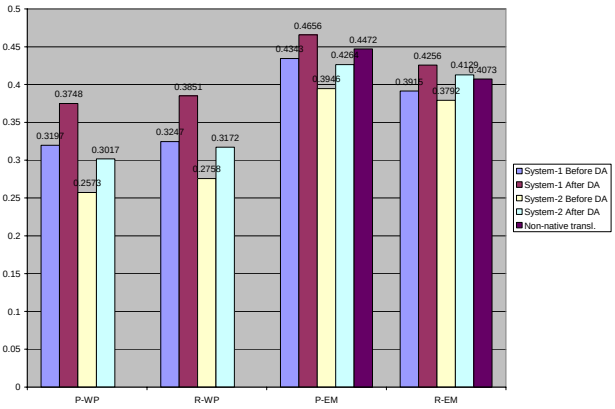

Figure 3. LTV scores – precision and recall

	P-WP	**R-WP**	**F-WP**
System-1 Before DA	0.3197	0.3247	0.3222
System-1 After DA	0.3748	0.3851	0.3799
System-2 Before DA	0.2573	0.2758	0.2663
System-2 After DA	0.3017	0.3172	0.3093
	P-EM	**R-EM**	**F-EM**
System-1 Before DA	0.4343	0.3915	0.4118
System-1 After DA	0.4656	0.4256	0.4447
System-2 Before DA	0.3946	0.3792	0.3868
System-2 After DA	0.4264	0.4129	0.4196
Non-native transl.	0.4472	0.4073	0.4263

Table 4. LTV scores

BLEU and LTV agree with human judgments with respect to ranking the two systems, although they differ in their precise scores. Results after dictionary update are better than before the update, and scores for System 1 are somewhat higher than for System 2; however, System 2 is shown to be capable of reaching System 1's baseline quality (the quality "before update") after its dictionary has been updated. The ratios of improvement and ratios of differences between systems are close to the ratios for human evaluation. This is an indication that human intuitive judgments about fluency, adequacy and usability of MT quality across systems and before and after the dictionary update are confirmed by the objective criteria: precision of N-gram matches in MT and the "gold standard" translation.

An important difference between the two automated metrics and the human evaluation results is the score for the non-native translation: BLEU seriously underestimates the quality of the human translation, LTV slightly less so. The explanation for this fact could be that for knowledge-based MT and for native-speaker human translations there is a close match between the adequacy and fluency of translation, but this is not the case for non-native translation (as well as for the output of statistical MT

systems, see (Babych, Hartley, Atwell, 2003)). Therefore, the N-gram precision is not a good model for usability of Non-native human translations, which doesn't use similar words that are required in "natural" English, and doesn't sufficiently match the N-grams in the "gold standard" translation, but nevertheless "makes sense" for the readers of the text. The second aspect of the explanation could be that BLEU is a much better measure for fluency than for adequacy; usability of emails, supposedly, has stronger links with the latter than with the former.

BLEU and LTV also indicate that emails are easier for MT than the whitepaper text: the absolute evaluation scores of both automated methods are higher for the emails.

LTV measures both precision and recall, so we may see that the recall measure is more stable across "easy" and "hard" texts, while precision changes much more if the type of the text changes. "Harder" texts, such as the whitepaper legal documents usually cause much greater over-generation of N-grams, but the under-generation of N-grams changes to a much smaller extent.

Correlation between automatic and human evaluation scores

Table 5 summarises correlation between automatic scores – BLEU and LTV and the human evaluation scores. The LTV and BLEU scores which previously have been found to closely correlate with corresponding human evaluation measures are underlined.

	LTV-P-WP	LTV-R-WP	LTV-F-WP
cFLU	0.984809	<u>0.989558</u>	0.988328
cADE	0.949595	<u>0.970463</u>	0.960599
	LTV-P-EM	LTV-R-EM	LTV-F-EM
cUSL/MT	0.905698	0.967349	<u>0.969011</u>
cUSL/MT+HT	0.593061	0.475047	0.562204
	BLEU-WP	BLEU-EM	
cFLU	<u>0.982683</u>		
cADE	0.945306		
cUSL/MT		0.933908	
cUSL/MT+HT		0.333796	

Table 5. Correlation between automatic and human evaluation scores

The chart shows that although BLEU provides scores which correlate closely with human judgments, especially for fluency, LTV outperforms BLEU for all the measured scores. The greatest advantage of the LTV is for adequacy and usability. Usability scores were not part of previous experiments, but the closest match for it is the LTV F-score, and the LTV Recall comes very close behind it.

The following conclusions can be drawn from the experiment:

1. Both automatic methods capture quality increase after dictionary update and rank systems correctly, in line with human judgments about MT quality.

2. The LTV method measures both precision and recall of N-gram matches, which allows flexible evaluation of different aspects of MT quality, such as adequacy and usability.

3. The usability metric integrates elements of adequacy and fluency, as is reflected in both human and automatic evaluation scores.

References

Akiba Y., K Imamura and E. Sumita. 2001. Using multiple edit distances to automatically rank machine translation output. In Proc. MT Summit VIII. p. 15–20.

Babych, B., Hartley, A., Atwell, E. 2003. Statistical Modelling of MT output corpora for Information Extraction. In: Proceedings of the Corpus Linguistics 2003 conference, Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds). Lancaster University (UK), 28-31 March 2003. pp. 62-70.

Babych, B. 2004. Weighted N-gram model for evaluating Machine Translation output. In: Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics. University of Birmingham, 6-7 January, 2004. pp. 15-22.

Brew C., Thompson H. 1994. Automatic Evaluation of Computer Generated Text. ARPA/ISTO Workshop on Human Language Technology, 1994. pp. 104-109.

Papineni, K., Roukos S, Ward T, Zhu W-J. 2002. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

Rajman, M. and T. Hartley. 2001. Automatically predicting MT systems ranking compatible with fluency, adequacy and Informativeness scores. Proceedings of the 4th ISLE Workshop on MT Evaluation, MT Summit VIII. Santiago de Compostela, September 2001. pp. 29-34.