

Extending BLEU MT Evaluation Method with Frequency Weighting

Bogdan Babych

Centre for Translation Studies
University of Leeds
Leeds, LS6 9JT, UK
bogdan@comp.leeds.ac.uk

Anthony Hartley

Centre for Translation Studies
University of Leeds
Leeds, LS6 9JT, UK
a.hartley@leeds.ac.uk

Abstract

We present the results of an experiment on extending the automatic method of Machine Translation evaluation BLEU with statistical weights for lexical items, such as tf.idf scores. We show that this extension gives additional information about evaluated texts; in particular it allows us to measure translation Adequacy, which, for statistical MT systems, is often overestimated by the baseline BLEU method. The proposed model uses a single human reference translation, which increases the usability of the proposed method for practical purposes. The model suggests a linguistic interpretation which relates frequency weights and human intuition about translation Adequacy and Fluency.

1. Introduction

Automatic methods for evaluating different aspects of MT quality – such as Adequacy, Fluency and Informativeness – provide an alternative to an expensive and time-consuming process of human MT evaluation. They are intended to yield scores that correlate with human judgments of translation quality and enable systems (machine or human) to be ranked on this basis. Several such automatic methods have been proposed in recent years. Some of them use human reference translations, e.g., the BLEU method (Papineni et al., 2002), which is based on comparison of N-gram models in MT and in a set of human reference translations.

However, a serious problem for the BLEU method is the lack of a model for relative importance of matched and mismatched items. Words in text usually carry an unequal informational load, and as a result are of differing importance for translation. It is reasonable to expect that the choices of right translation equivalents for certain key items, such as expressions denoting principal events, event participants and relations in a text are more important in the eyes of human evaluators than choices of functional words and a syntactic perspective for sentences. Accurate rendering of these key items by an MT system boosts the quality of translation. Therefore at least for evaluation of translation Adequacy the proper choice of translation equivalents for important pieces of information should count more than the choice of words without clear translation equivalent in the source text, which are used for linking purposes. (The later may be more important for Fluency evaluation).

The problem of different significance of N-gram matches is related to the issue of legitimate variation in human translations (e.g., functional words and different morphological forms tend to vary much more across independently produced human translations than other words). BLEU accounts for legitimate translation variation by using a set of several human reference translations, which are believed to be representative of several equally acceptable ways of translating any source segment. This is motivated by the need not to penalise deviations from the set of N-grams in single reference, although the requirement of multiple human references makes automatic evaluation more expensive.

However, the “significance” problem is not directly addressed by the BLEU method. On the one hand, the matched items that are present in

several human references receive the same weights as items found in just one of the references. On the other hand the model of legitimate translation variation cannot fully accommodate the issue of varying degrees of “salience” for matched lexical items, since alternative synonymic translation equivalents may also be highly significant for an adequate translation (from the human perspective). Therefore it is reasonable to suggest that introduction of a model which approximates intuitions about the significance of the matched N-grams will improve the correlation between automatically computed MT evaluation scores and human evaluation scores for translation Adequacy.

In this paper we present the result of an experiment on augmenting BLEU N-gram comparison with statistical weight coefficients which capture the degree of word’s salience within a given document: the standard tf.idf measure used in the vector-space model for Information Retrieval (Salton and Leck, 1968) and the S-score proposed for evaluating MT output corpora for the purposes of Information Extraction (Babych et al., 2003). Both scores are computed for each term in each of the 100 human reference translations from French into English available in DARPA-94 MT evaluation corpus (White et al., 1994).

The proposed weighted N-gram model for MT evaluation is tested on a set of translations by four different MT systems available in the DARPA corpus, and is compared with the results of the baseline BLEU method with respect to their correlation with human evaluation scores.

The scores produced by the N-gram model with tf.idf and S-Score weights are shown to be consistent with baseline BLEU evaluation results for Fluency and outperform the BLEU scores for Adequacy (correlation for the S-score weighting is slightly higher). We also show that the weighted model may be still reliably used if there is only one human reference translation for an evaluated text.

Besides saving cost, the ability to dependably work with a single human translation has an additional advantage: it is now possible to create Recall-based evaluation measures for MT, which has been problematic for evaluation with multiple reference translations, since only one of the choices from the reference set is used in transla-

tion (Papineni et al. 2002:314). Notably, Recall of weighted N-grams is found to be a good estimation of human judgements about translation Adequacy. Using *weighted* N-grams is essential for predicting adequacy, since correlation of Recall for non-weighted N-grams is much lower.

It is possible that other automatic methods which use human translations as a reference may also benefit from an introduction of an explicit model for term significance, since so far these methods also implicitly assume that all words are equally important in human translation, and use all of them, e.g., for measuring edit distances (Akiba et al, 2001; 2003).

The weighted N-gram model has been implemented as an MT evaluation toolkit (which includes a Perl script, example files and documentation). It computes evaluation scores with tf.idf and S-score weights for translation Adequacy and Fluency.

2. Set-up of the experiment

The experiment used French–English translations available in the DARPA-94 MT evaluation corpus. The corpus contains 100 French texts translated into English by 4 different MT systems: “Systran”, “Globalink”, “Metal” and “Candide”. It also contains 2 independent human translations of each text and evaluation scores given by human evaluators to each of the 400 machine translations. Human scores were given for 3 parameters of translation quality: “Adequacy”, “Fluency” and “Informativeness”.

In the first stage of the experiment, each of the two sets of human translations was used to compute tf.idf and S-scores for each word in each of the 100 texts. The tf.idf score was calculated as:

$$tf.idf(i,j) = (1 + \log (tf_{i,j})) \log (N / df_i),$$

if $tf_{i,j} \geq 1$; where:

- $tf_{i,j}$ is the number of occurrences of the word w_i in the document d_j ;
- df_i is the number of documents in the corpus where the word w_i occurs;
- N is the total number of documents in the corpus.

The S-score was calculated as:

$$S(i, j) = \log \frac{(P_{doc(i,j)} - P_{corp-doc(i)}) \times (N - df_{(i)}) / N}{P_{corp(i)}}$$

where:

- $P_{doc(i,j)}$ is the relative frequency of the word in the text; (“Relative frequency” is the number of tokens of this word-type divided by the total number of tokens).
- $P_{corp-doc(i)}$ is the relative frequency of the same word in the rest of the corpus, without this text;
- $(N - df_{(i)}) / N$ is the proportion of texts in the corpus, where this word does not occur (number of texts, where it is not found, divided by number of texts in the corpus);
- $P_{corp(i)}$ is the relative frequency of the word in the whole corpus, including this particular text.

In the second stage we carried out N-gram based MT evaluation, measuring Precision and Recall of N-grams in MT output using a single human reference translation. N-gram counts were adjusted with the tf.idf weights and S-scores for every matched word.

The following procedure was used to integrate the S-scores into N-gram counts:

- If for a lexical item in a text the *S-score* > 1, all counts for the N-grams containing this item are increased by the *S-score* (not just by 1, as in the baseline BLEU approach).

- If the *S-score* ≤ 1; the usual N-gram count is applied: the number is increased by 1.

The original matches used for BLEU and the weighted matches are both calculated. The following additions have been made to the Perl script of the BLEU tool: apart from the operator which increases counts for every matched N-gram \$ngr by 1, i.e.:

```
$ngr .= $words[$i+$j] . " ";
$$hashNgr{$ngr}++;
```

the following code was introduced:

```
[...]
$WORD = $words[$i+$j];
$WEIGHT = 1;
if(exists
    $WordWeight{$TtxtN}{$WORD}&&
    $WordWeight{$TtxtN}{$WORD} > 1){
    $WEIGHT=
        $WordWeight{$TtxtN}{$WORD};
}

$ngr .= $words[$i+$j] . " ";
$$hashNgr{$ngr}++;

$$hashNgrWEIGHTED{$ngr}+= $WEIGHT;
[...]
```

– where the hash data structure:

```
$WordWeight{$TtxtN}{$WORD}=$WEIGHT
```

represents the table of tf.idf scores or S-scores for words in every text in the corpus.

The weighted N-gram evaluation scores of Precision, Recall and F-measure may be produced for a segment, for a text or for a corpus of translations generated by an MT system.

In the third stage of the experiment the weighted Precision and Recall scores were tested for correlation with human scores for the same texts and compared to the results of similar tests for standard BLEU evaluation.

Finally we addressed the question whether the proposed MT evaluation method allows us to use a single human reference translation reliably. In order to assess the stability of the weighted evaluation scores with a single reference, two runs of the experiment were carried out. The first run used the “Reference” human translation, while the second run used the “Expert” human translation (each time a single reference translation was used). The scores for both runs were compared using a standard deviation measure.

3. The results of the MT evaluation with frequency weights

With respect to evaluating MT systems, the correlation for the weighted N-gram model was found to be stronger, for both Adequacy and Fluency, the improvement being highest for Adequacy. These results are due to the fact that the weighted N-gram model gives much more accurate predictions about the statistical MT system “Candide”, whereas the standard BLEU approach tends to over-estimate its performance for translation Adequacy.

Table 1 summarises the evaluation scores for BLEU as compared to tf.idf weighted scores. It shows the following figures:

- Human evaluation scores for Adequacy and Fluency (the mean scores for all texts produced by each MT system);
- BLEU scores produced using 2 human reference translations and the default script settings (N-gram size = 4);
- Precision, Recall and F-score for the weighted N-gram model produced with 1 human reference translation and N-gram size = 4.
- Pearson’s correlation coefficient r for Precision, Recall and F-score correlated with human scores for Adequacy and Fluency $r(2)$ (with 2

degrees of freedom) for the sets which include scores for the 4 MT systems.

The scores at the top of each cell show the results for the first run of the experiment, which used the “Reference” human translation; the scores at the bottom of the cells represent the results for the second run with the “Expert” human translation.

System [ade] / [flu]	BLEU [1&2]	Prec. (w) 1/2	Recall (w) 1/2	Fscore (w) 1/2
CANDIDE 0.677 / 0.455	0.3561	0.4767 0.4709	0.3363 0.3324	0.3944 0.3897
GLOBALINK 0.710 / 0.381	0.3199	0.4289 0.4277	0.3146 0.3144	0.3630 0.3624
MS 0.718 / 0.382	0.3003	0.4217 0.4218	0.3332 0.3354	0.3723 0.3737
REVERSO NA / NA	0.3823	0.4760 0.4756	0.3643 0.3653	0.4127 0.4132
SYSTRAN 0.789 / 0.508	0.4002	0.4864 0.4813	0.3759 0.3734	0.4241 0.4206
Corr $r(2)$ with [ade] – MT	0.5918	0.3399 0.3602	0.7966 0.8306	0.6479 0.6935
Corr $r(2)$ with [flu] – MT	0.9807	0.9665 0.9721	0.8980 0.8505	0.9853 0.9699

Table 1. BLEU vs tf.idf weighted scores.

Table 2 summarises the same scores for BLEU as compared to S-score weighed evaluation.

System [ade] / [flu]	BLEU [1&2]	Prec. (w) 1/2	Recall (w) 1/2	Fscore (w) 1/2
CANDIDE 0.677 / 0.455	0.3561	0.4570 0.4524	0.3281 0.3254	0.3820 0.3785
GLOBALINK 0.710 / 0.381	0.3199	0.4054 0.4036	0.3086 0.3086	0.3504 0.3497
MS 0.718 / 0.382	0.3003	0.3963 0.3969	0.3237 0.3259	0.3563 0.3579
REVERSO NA / NA	0.3823	0.4547 0.4540	0.3563 0.3574	0.3996 0.4000
SYSTRAN 0.789 / 0.508	0.4002	0.4633 0.4585	0.3666 0.3644	0.4094 0.4061
Corr $r(2)$ with [ade] – MT	0.5918	0.2945 0.2996	0.8046 0.8317	0.6184 0.6492
Corr $r(2)$ with [flu] – MT	0.9807	0.9525 0.9555	0.9093 0.8722	0.9942 0.9860

Table 2. BLEU vs S-score weights.

It can be seen from the table that there is a strong positive correlation between the baseline BLEU scores and human scores for Fluency: $r(2)=0.9807, p < 0.05$. However, the correlation with Adequacy is much weaker and is not statistically significant: $r(2)= 0.5918, p > 0.05$. The

most serious problem for BLEU is predicting scores for the statistical MT system “Candide”, which was judged to produce relatively fluent, but largely inadequate translation. For other MT systems (developed with the knowledge-based MT architecture) the scores for Adequacy and Fluency are consistent with each other: more fluent translations are also more adequate. BLEU scores go in line with “Candide’s” Fluency scores, and do not account for its Adequacy scores. When “Candide” is excluded from the evaluation set, r correlation goes up, but it is still lower than the correlation for Fluency and remains statistically insignificant: $r(1)=0.9608, p > 0.05$. Therefore, the baseline BLEU approach fails to consistently predict scores for Adequacy.

The proposed weighted N-gram model outperforms BLEU in its ability to predict Adequacy scores: weighted Recall scores have much stronger correlation with Adequacy (which is still statistically insignificant for MT-only evaluation. Correlation figures for S-score-based weights are slightly higher than for tf.idf weights (S-score: $r(2)= 0.8046, p > 0.05$; $r(2)= 0.8317, p > 0.05$, tf.idf score: $r(2)= 0.7966, p > 0.05$; $r(2)= 0.8306, p > 0.05$).

This is achieved by reducing overestimation for the “Candide” system, moving its scores closer to human judgements about its quality in this respect. However, this is not completely achieved: “Candide” is still slightly better than Globalink and MS according to the weighted Recall score, but it is slightly worse than Globalink according to human judgements about Adequacy.

For both methods – BLEU and the Weighted N-gram evaluation – Adequacy is found harder to predict than Fluency. This is due to the fact that there is no good linguistic model of translation adequacy which can be easily formalised. The introduction of S-score weights may be a useful step towards developing such a model, since correlation scores with Adequacy are much better for the Weighted N-gram approach than for BLEU.

Also from the linguistic point of view, S-score weights and N-grams may only be reasonably good approximations of Adequacy, which involves a wide range of factors, like syntactic and semantic issues that cannot be captured by N-gram matches and require a thesaurus and other knowledge-based extensions. Accurate formal

models of translation transformations may also be useful for improving automatic evaluation of Adequacy.

The proposed evaluation method also preserves the ability of BLEU to consistently predict scores for Fluency: weighted F-scores have a strong positive correlation with this aspect of MT quality, the figures are very similar to the values for BLEU (*S-score*: $r(2) = 0.9942$, $p < 0.01$; $r(2) = 0.9860$, $p < 0.01$; *tf.idf score*: $r(2) = 0.9853$, $p < 0.01$; $r(2) = 0.9699$, $p < 0.05$).

However, strong correlation with Fluency in the proposed method is achieved by different means than in BLEU: instead of the “modified N-gram precision” suggested in (Papineni et al., 2002:312) and 2 human reference translations, we used the combined F-score (Precision and Recall weighted equally) and only 1 human reference translation. Counts of weighted N-grams were straightforward; no modifications to standard Precision and Recall measures have been applied.

These two major methodological differences are linked with the previous point (strong correlation between the weighted N-gram Recall and Adequacy): using 1 human reference with uniform results means that there is no more “trouble with Recall” (Papineni et al., 2002:314) – system’s ability to avoid under-generation of N-grams can now be reliably measured. Therefore, it became also possible to compute the F-score. As a result calculations became much simpler: Recall was found to give good estimation for Adequacy, and the F-score reliably predicts Fluency.

Certainly, using a single human reference translation instead of multiple translations will increase usability of N-gram based MT evaluation tools.

Moreover, this suggests a new linguistic interpretation of the nature of these two quality criteria: it is intuitively plausible that Fluency subsumes, i.e. presupposes Adequacy (similarly to the way the F-score subsumes Recall). The F-score correlates stronger with Adequacy than both of its components: Precision and Recall; similarly Adequacy might make a contribution to Fluency together with some other factors. It is conceivable that people need adequate translations (or at least translations that make sense) in

order to be able to make judgments about naturalness, or fluency.

Being able to make some sense out of a text could be the major ground for judging Adequacy: sensible mistranslations in MT are relatively rare events. This may be the consequence of a principle similar to the “second law of thermodynamics” applied to text structure, – it is much harder to create some alternative sense than to destroy the existing sense in translation, so the majority of inadequate translations are just nonsense. However, fluent mistranslations are even rarer than disfluent ones, according to the same principle. A real difference in scores is made by segments which make sense and may or may not be fluent, and things which do not make any sense and about which it is hard to tell whether they are fluent.

This suggestion may be empirically tested: if Adequacy is a necessary precondition for Fluency, there should be a greater inter-annotator disagreement in Fluency scores on texts or segments which have lower Adequacy scores. Empirical assessment of this hypothesis will be a topic of future research.

We note that for the DARPA corpus the correlation scores presented are highest if the evaluation unit is an entire corpus of translations produced by an MT system, and for text-level evaluation, correlation is much lower. The similar observation was made in (Papineni et al., 2002: 313). This may be due to the fact that human judges are less consistent, especially for puzzling segments that do not fit the scoring guidelines, like nonsense segments for which it is hard to decide whether they are fluent or even adequate. However, this randomness is leveled out if the evaluation unit increases in size – from the text level to the corpus level.

Automatic evaluation methods such as BLEU (Papineni et al., 2002), RED (Akiba et al., 2001), or weighted N-gram model proposed here – may be more consistent in judging quality as compared to human evaluators, but human judgments remain the only criteria for meta-evaluating the automatic methods.

4. Stability of weighted evaluation scores

In this section we investigate how reliable is the use of a single human reference translation. The

stability of the scores is central to the issue of computing Recall and reducing the cost of automatic evaluation.

We also would like to compare stability of our results with stability of the baseline N-gram model with a single reference. In order to carry out this comparison we re-implemented simplified BLEU-type system which produced baseline counts of N-grams without weights and used a single human reference translation at a time. This system works exactly as the weighted N-gram model, but instead of using S-score weights, it just counts the N-grams.

In this stage of the experiment we measured the changes that occur for the scores of MT systems if an alternative reference translation is used – both for the baseline N-gram counts and for the weighted N-gram model. Standard deviation was computed for each pair of evaluation scores produced by the two runs of the system with alternative human references. An average of these standard deviations is the measure of stability for a given score. The results of these calculations are presented in Table 3.

	systems	StDev-basln	StDev-tf.idf	StDev-S-score
P	candide	0.004	0.0041	0.0033
	globalink	0.0011	0.0008	0.0013
	ms	0.0002	0.0001	0.0004
	reverso	0.0018	0.0003	0.0005
	systran	0.0034	0.0036	0.0034
	AVE SDEV	0.0021	0.0018	0.0018
R	candide	0.0011	0.0028	0.0019
	globalink	0.0013	0.0001	0
	ms	0.0023	0.0016	0.0016
	reverso	0.0009	0.0007	0.0008
	systran	0.0008	0.0018	0.0016
	AVE SDEV	0.0013	0.0014	0.0012
F	candide	0.0025	0.0033	0.0025
	globalink	0.0001	0.0004	0.0005
	ms	0.0009	0.001	0.0011
	reverso	0.0005	0.0004	0.0003
	systran	0.0021	0.0025	0.0023
	AVE SDEV	0.0012	0.0015	0.0013

Table 3. Stability of scores

Both the baseline and the weighted N-gram approaches give relatively stable results: the standard deviation was not greater than 0.0021,

which means that both will produce reliable figures with just a single human reference translation (although interpretation of the score with a single reference should be different than with multiple references).

This comparison also shows that the absence of significant difference in standard deviation figures between the baseline and the weighted N-gram counts confirms our initial suggestion that the model of term’s importance for translation cannot be straightforwardly derived from the model of the legitimate translation variation, used in BLEU.

5. Conclusion and future work

The results for weighted N-gram models have a significantly higher correlation with human intuitive judgements about translation Adequacy and Fluency than the baseline N-gram evaluation measures which are used in the BLEU MT evaluation toolkit. This shows that they are a promising direction of research in the field of automatic MT evaluation. Future work will involve applying our approach to evaluating MT into languages other than English.

However, the results of the experiment may also have implications for MT development: significance weights may be used to rank the relative “importance” of translation equivalents. At present all MT architectures (knowledge-based, example-based, and statistical) treat all translation equivalents equally, so MT systems cannot dynamically prioritise rule applications, and translations of the central concepts in texts are often lost among literal translations of function words, lexical collocations not appropriate for the target language, etc. For example, for statistical MT significance weights of lexical items may indicate which words have to be introduced into the target text using the *translation model* for source and target languages, and which need to be brought there by the *language model* for the target corpora. Similar ideas may be useful for the Example-based and Rule-based MT architectures. The general idea is that not everything in the source text needs to be translated, and the significance weights schedule the priority for application of translation equivalents and may motivate the need for application of compensation strategies in translation. They allow us to

make an approximate distinction between salient words which require proper translation equivalents and linking material both in the source and in the target texts. Exploring applicability of this idea to various MT architectures is a direction for future research.

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.
- Akiba Y., E. Sumita, H. Nakaiwa, S. Yamamoto and H.G. Okuno. 2003. Experimental Comparison of MT Evaluation Methods: RED vs. BLEU. In *Proc. MT Summit IX*, URL: <http://www.amtaweb.org/summit/MTSummit/FinalPapers/55-Akiba-final.pdf>.
- Babych, B; Hartley, A.; Atwell, E. Statistical Modelling of MT output corpora for Information Extraction. In: Proceedings of the Corpus Linguistics 2003 conference, edited by Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery. Lancaster University (UK), 28 - 31 March 2003. Pp. 62-70.
- 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.
- Salton, G. and M.E. Lesk. 1968. Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1), 8-36.
- White, J., T. O'Connell and F. O'Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons and future approaches. *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*. Columbia, MD, October 1994. pp. 193-205.