Extending the BLEU MT Evaluation Method with Frequency Weightings

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

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

Abstract

We present the results of an experiment on extending the automatic method of Machine Translation evaluation BLUE 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 output 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 then choices of function 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 (Fidelity), the proper choice of translation equivalents for important pieces of information should count more than the choice of words which are used for structural purposes and without a clear translation equivalent in the source text. (The latter 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, when certain words are less stable than others across independently produced human translations. 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 a 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 (Babych and Hartley, 2004). 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 a 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 (where the correlation for the S-score weighting is higher). We also show that the weighted model may still be 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 translation (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. The toolkit is available at http://www.comp.leeds.ac.uk/bogdan/evalMT.html

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 news texts (each text is about 350 words long) translated into English by 5 different MT systems: "Systran", "Reverso", "Globalink", "Metal", "Candide" and scored by human evaluators; there are no human scores for "Reverso", which was added to the corpus on a later stage. The corpus also contains 2 independent human translations of each text. Human evaluation scores are available for each of the 400 texts translated by the 4 MT systems for 3 parameters of translation quality: "Adequacy", "Fluency" and "Informativeness". The *Adequacy* (Fidelity) scores are given on a 5-point scale by comparing MT with a human reference translation. The Adequacy parameter captures how much of the original content of a text is conveyed, regardless of how grammatically imperfect the output might be. The *Fluency* scores (also given on a 5-point scale) determine intelligibility of MT without reference to the source text, i.e., how grammatical and stylistically natural the translation appears to be. The *Informativeness* scores (which we didn't use for our experiment) determine whether there is enough information in MT output to enable evaluators to answer multiple-choice questions on its content (White, 2003:237)

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:

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

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).
- $P_{corp\text{-}doc(i)}$ 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 / tf.idf scores for a lexical item into N-gram counts. For every word in a given text which received an S-score and tf.idf score on the basis of the human reference corpus, all counts for the N-grams containing this word are increased by the value of the respective score (not just by 1, as in the baseline BLEU approach).

The original matches used for BLEU and the weighted matches are both calculated. The following changes 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 = 0;
if(exists
  $WordWeight{$TxtN}{$WORD}){
    $WEIGHT=
     $WordWeight{$TxtN}{$WORD};
}

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

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

– where the hash data structure:

```
 $WordWeight{$TxtN}{$WORD}=$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 present the baseline results for non-weighted Precision, Recall and F-score. 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.** **1/2**	**Recall** **1/2**	**Fscore** **1/2**
CANDIDE 0.677 / 0.455	0.3561	0.4068 0.4012	0.3806 0.3790	0.3933 0.3898
GLOBALINK 0.710 / 0.381	0.3199	0.3429 0.3414	0.3465 0.3484	0.3447 0.3449
MS 0.718 / 0.382	0.3003	0.3289 0.3286	0.3650 0.3682	0.3460 0.3473
REVERSO *NA / NA*	0.3823	0.3948 0.3923	0.4012 0.4025	0.3980 0.3973
SYSTRAN 0.789 / 0.508	0.4002	0.4029 0.3981	0.4129 0.4118	0.4078 0.4049
Corr ***r(2)*** *with* ***[ade]*** *– MT*	0.5918	0.1809 0.1871	0.6691 0.6988	0.4063 0.4270
Corr ***r(2)*** *with* ***[flu]*** *– MT*	0.9807	0.9096 0.9124	0.9540 0.9353	**0.9836** **0.9869**

Table 1. Baseline non-weighted scores.

Table 2 summarises the evaluation scores for BLEU as compared to tf.idf weighted scores, and Table 3 summarises the same scores 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	0.5242 0.5176	0.3094 0.3051	0.3892 0.3839
GLOBALINK 0.710 / 0.381	0.3199	0.4905 0.4890	0.2919 0.2911	0.3660 0.3650
MS 0.718 / 0.382	0.3003	0.4919 0.4902	0.3083 0.3100	0.3791 0.3798
REVERSO *NA / NA*	0.3823	0.5336 0.5342	0.3400 0.3413	0.4154 0.4165
SYSTRAN 0.789 / 0.508	0.4002	0.5442 0.5375	0.3521 0.3491	0.4276 0.4233
Corr ***r(2)*** *with* ***[ade]*** *– MT*	0.5918	0.5248 0.5561	0.8354 0.8667	0.7691 0.8119
Corr ***r(2)*** *with* ***[flu]*** *– MT*	0.9807	**0.9987** **0.9998**	0.8849 0.8350	0.9408 0.9070

Table 2. BLEU *vs* tf.idf weighted scores.

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	0.5034 0.4982	0.2553 0.2554	0.3388 0.3377
GLOBALINK 0.710 / 0.381	0.3199	0.4677 0.4672	0.2464 0.2493	0.3228 0.3252
MS 0.718 / 0.382	0.3003	0.4766 0.4793	0.2635 0.2679	0.3394 0.3437
REVERSO *NA / NA*	0.3823	0.5204 0.5214	0.2930 0.2967	0.3749 0.3782
SYSTRAN 0.789 / 0.508	0.4002	0.5314 0.5218	0.3034 0.3022	0.3863 0.3828
Corr ***r(2)*** *with* ***[ade]*** *– MT*	0.5918	0.6055 0.6137	**0.9069** **0.9215**	0.8574 0.8792
Corr ***r(2)*** *with* ***[flu]*** *– MT*	0.9807	0.9912 0.9769	0.8022 0.7499	0.8715 0.8247

Table 3. 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, but 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.

Correlation figures between non-weighted N-gram counts and human scores are similar to the results for BLEU: the highest and statistically significant correlation is between the F-score and Fluency: *r(2)=0.9836, p<0.05, r(2)=0.9869, p<0.01*, and there is somewhat smaller and statistically significant correlation with Precision. This confirms the need to use *modified* Precision in the BLEU method that also in certain respect integrates Recall.

The proposed weighted N-gram model outperforms BLEU and non-weighted N-gram evaluation in its ability to predict Adequacy scores: weighted Recall scores have much stronger correlation with Adequacy (which for MT-only evaluation is still statistically insignificant at the level *p<0.05*, but come very close to that point: *t=3.729* and *t=4.108*; the required value for *p<0.05* is *t=4.303*).

Correlation figures for S-score-based weights are higher than for tf.idf weights (*S-score*: *r(2)= 0.9069, p > 0.05*; *r(2)= 0.9215, p > 0.05, tf.idf score*: *r(2)= 0.8354, p >0.05; r(2)= 0.8667, p >0.05*).

The improvement in the accuracy of evaluation for the weighted N-gram model can be illustrated by the following example of translating the French sentence:

ORI-French: Les trente-huit chefs d'entreprise mis en examen dans le dossier ont déjà fait l'objet d'auditions, mais trois d'entre eux ont été confrontés, mercredi, dans la foulée de la confrontation "politique".

English translations of this sentence by the knowledge-based system Systran and statistical MT system Candide have an equal number of matched unigrams (highlighted in italic), therefore conventional unigram Precision and Recall scores are the same for both systems. However, for each translation two of the matched unigrams are different (underlined) and receive different frequency weights (shown in brackets):

MT “Systran”:

The thirty-eight *heads* (tf.idf=**4.605**; S=**4.614**) *of* undertaking put in examination *in the* file *already* were the subject of hearings, *but three of them were* confronted, *Wednesday*, in *the* tread of "*political*" *confrontation* (tf.idf=**5.937**; S=**3.890**).

Human translation “Expert”:

The thirty-eight *heads* ***of*** companies questioned ***in the* case had *already*** been heard, ***but three of them were*** brought together ***Wednesday*** following ***the*** "***political***" *confrontation*.

MT “Candide”:

The thirty-eight counts *of* company put into consideration *in the* *case* (tf.idf=**3.719**; S=**2.199**) *already* *had* (tf.idf=**0.562**; S=**0.000**) the object of hearings, *but three of them were* checked, *Wednesday*, in *the* path of confrontal "*political*."

(In the human translation the unigrams matched by the Systran output sentence are in italic, those matched by the Candide sentence are in bold).

It can be seen from this example that the unigrams matched by Systran have higher term frequency weights (both tf.idf and S-scores):

heads (tf.idf=**4.605**;S=**4.614**)
confrontation (tf.idf=**5.937**;S=**3.890**)

The output sentence of Candide instead matched less salient unigrams:

case (tf.idf=**3.719**;S=**2.199**)
had (tf.idf=**0.562**;S=**0.000**)

Therefore for the given sentence weighted unigram Recall (i.e., the ability to avoid undergeneration of salient unigrams) is higher for Systran than for Candide (Table 4):

	Systran	Candide
R	0.6538	0.6538
R * tf.idf	0.5332	0.4211
R * S-score	0.5517	0.3697
P	0.5484	0.5484
P * tf.idf	0.7402	0.9277
P * S-score	0.7166	0.9573

Table 4. Recall, Precision, and weighted scores

Weighted Recall scores capture the intuition that the translation generated by Systran is more adequate than the one generated by Candide, since it preserves more important pieces of information.

On the other hand, weighted Precision scores are higher for Candide. This is due to the fact that Systran over-generates (doesn’t match in the human translation) much more “exotic”, unordinary words, which on average have higher cumulative salience scores, e.g., `undertaking`, `examination`, `confronted`, `tread` – vs. the corresponding words "over-generated" by Candide: `company`, `consideration`, `checked`, `path`. In some respect higher weighted precision can be interpreted as higher Fluency of the Candide's output sentence, which intuitively is perceived as sounding more naturally (although not making much sense).

On the level of corpus statistics the weighted Recall scores go in line with Adequacy, and weighted Precision scores (as well as the Precision-based BLEU scores) – with Fluency, which confirms such interpretation of weighted Precision and Recall scores in the example above. On the other hand, Precision-based scores and non-weighted Recall scores fail to capture Adequacy.

The improvement in correlation for weighted Recall scores with Adequacy 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: although in terms of Recall weighted by the S-scores Candide is correctly ranked below MS (and not ahead of it, as with the BLEU scores), it is still slightly ahead of Globalink, contrary to human evaluation results.

For both methods – BLEU and the Weighted N-gram evaluation – Adequacy is found to be 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 variation 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: Precision weighted by tf.idf scores has the strongest positive correlation with this aspect of MT quality, which is slightly better than the values for BLEU; (*S-score: $r(2)= 0.9912$, $p<0.01$; $r(2)= 0.9769$, $p<0.05$; tf.idf score: $r(2)= 0.9987$, $p<0.001$; $r(2)= 0.9998$, $p<0.001$*).

The results suggest that weighted Precision gives a good approximation of Fluency. Similar results with non-weighted approach are only achieved if some aspect of Recall is integrated into the evaluation metric (either as *modified* precision, as in BLEU, or as an aspect of the F-score). Weighted Recall (especially with S-scores) gives a reasonably good approximation of Adequacy.

On the one hand using 1 human reference with uniform results is essential for our methodology, since it means that there is no more "trouble with Recall" (Papineni et al., 2002:314) – a system's ability to avoid under-generation of N-grams can now be reliably measured. On the other hand, using a single human reference translation instead of multiple translations will certainly increase the usability of N-gram based MT evaluation tools.

The fact that non-weighted F-scores also have high correlation with Fluency 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, which among all other scores gives the best correlation with Adequacy). The non-weighted F-score correlates more strongly with Fluency than either 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, – in practice it is much rarer to some alternative sense to be created (even if the number of possible error types could be significant), than to destroy the existing sense in translation, so the majority of inadequate translations are just nonsense. However, in contrast to human translation, fluent mistranslations in MT 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. This 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. A 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 the 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 the stability of our results with the stability of the baseline non-weighted N-gram model using a single reference.

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 5.

	systems	StDev-basln	StDev-tf.idf	StDev-S-score
P	candide	0.004	0.0047	0.0037
	globalink	0.0011	0.0011	0.0004
	ms	0.0002	0.0012	0.0019
	reverso	0.0018	0.0004	0.0007
	systran	0.0034	0.0047	0.0068
	AVE SDEV	0.0021	0.0024	0.0027
R	candide	0.0011	0.003	0.0001
	globalink	0.0013	0.0006	0.0021
	ms	0.0023	0.0012	0.0031
	reverso	0.0009	0.0009	0.0026
	systran	0.0008	0.0021	0.0008
	AVE SDEV	0.0013	0.0016	0.0017
F	candide	0.0025	0.0037	0.0008
	globalink	0.0001	0.0007	0.0017
	ms	0.0009	0.0005	0.003
	reverso	0.0005	0.0008	0.0023
	systran	0.0021	0.003	0.0025
	AVE SDEV	0.0012	0.0018	0.0021

Table 5. Stability of scores

Standard deviation for weighted scores is generally slightly higher, but both the baseline and the weighted N-gram approaches give relatively stable results: the average standard deviation was not greater than 0.0027, 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).

Somewhat higher standard deviation figures for the weighted N-gram model confirm the suggestion that a word's importance for translation cannot be straightforwardly derived from the model of the legitimate translation variation implemented in BLEU and needs the salience weights, such as tf.idf or S-scores.

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. Future work will apply our approach to evaluating MT into languages other than English, extending the experiment to a larger number of MT systems built on different architectures and to larger corpora.

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 excessively literal translations of less important concepts and function words. 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 different pieces of information expressed in the source text are not equally important for translation: MT systems that have no means for prioritising this information often introduce excessive information noise into the target text by literally translating structural information, etymology of proper names, collocations that are unacceptable in the target language, etc. This information noise often obscures important translation equivalents and prevents the users from focusing on the relevant bits. MT quality may benefit from filtering out this excessive information as much as from frequently recommended extension of knowledge sources for MT systems. The significance weights may schedule the priority for retrieving translation equivalents and motivate application of compensation strategies in translation, e.g., adding or deleting implicitly inferable information in the target text, using non-literal strategies, such as transposition or modulation (Vinay and Darbelnet, 1995). Such weights may allow MT systems to make an approximate distinction between salient words which require proper translation equivalents and structural material both in the source and in the target texts. Exploring applicability of this idea to various MT architectures is another direction for future research.

Acknowledgments

We are very grateful for the insightful comments of the three anonymous reviewers.

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., A. Hartley and E. Atwell. 2003. Statistical Modelling of MT output corpora for Information Extraction. In: *Proceedings of the Corpus Linguistics 2003 conference*, Lancaster University (UK), 28 - 31 March 2003, pp. 62-70.

Babych, B. and A. Hartley. 2004. Modelling legitimate translation variation for automatic evaluation of MT quality. In: *Proceedings of LREC 2004* (forthcoming).

Papineni, K., S. Roukos, T. Ward, W.-J. Zhu. 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.

Vinay, J.P. and J.Darbelnet. 1995. Comparative stylistics of French and English : a methodology for translation / translated and edited by Juan C. Sager, M.-J. Hamel. J. Benjamins Pub., Amsterdam, Philadelphia.

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.

White, J. 2003. How to evaluate machine translation. In: H. Somers. (Ed.) Computers and Translation: a translator's guide. Ed. J. Benjamins B.V., Amsterdam, Philadelphia, pp. 211-244.