

# Weighted N-gram model for evaluating Machine Translation output

**Bogdan Babych**

Centre for Translation Studies

University of Leeds, UK

Department of Computer Science

University of Sheffield, UK

[bogdan@comp.leeds.ac.uk](mailto:bogdan@comp.leeds.ac.uk)

## Abstract

I present the results of an experiment on extending an automatic method of Machine Translation evaluation (BLEU) with weights for the statistical significance of lexical items. I 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 also improves the stability of evaluation scores with a single human reference translation, which increases the usability of the proposed method for practical purposes. The model suggests linguistic a interpretation which develops deeper understanding of human intuitions 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., 2001), which is based on comparison of N-gram models.

However, a serious problem for the BLEU method is legitimate variation in human translations. BLEU tends to penalise any deviation from the reference set of N-grams, despite the fact that usually there are several equally acceptable ways of translating any source segment, and thus any deviations within MT output can be only partially attributed to errors. Usually multiple human reference translations for the same text are needed in order to solve this problem, which makes automatic evaluation more expensive.

I present the result of an experiment on developing an alternative solution to this problem: augmenting BLEU N-gram comparison with weight coefficients that characterise the potential stability of N-grams across different human translations. This idea is derived from the field of Information Extraction: some pieces of information in a text are more important for conveying its core meaning (e.g., names of events and event participants) and consequently need to be preserved in any translation; less central information (e.g., the choice of morpho-syntactic context and certain function words) is subject to greater variation since the translator has a greater freedom of choice.

Weights for N-grams are automatically generated on the basis of contrasting N-gram frequencies in a given text and in the rest of the corpus of texts to be translated. The ranking of N-grams so obtained is tested by comparing it to the

actual stability of N-grams in two human translations available for 100 texts in the DARPA-94 MT evaluation corpus (White et al., 1994). The resulting BLEU scores are adjusted by the N-gram stability weights: the contribution of matching N-grams to the MT evaluation score is proportional to the assigned weights.

The proposed approach to MT evaluation is tested on a set of translations from French into English by different MT systems available in the DARPA corpus, and is compared with the results of the baseline BLEU method.

The scores produced by the weighted N-gram model are shown to be consistent with baseline BLEU evaluation results for Fluency and outperform the BLEU scores for Adequacy. Moreover, they may be also reliably used if there is only one human reference translation for an evaluated text.

Besides saving cost, the ability to reliably use 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. 2001:4). However, 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.

The intuition behind the experiment is that words in text carry an unequal informational load, and as a result are of differing importance for translation. More informative (e.g., content) words tend to recur across different human translations. Accurate rendering of these words by an MT system boosts the quality of translation. However, automatic methods which use a human translation as a reference, implicitly assume that all words are equally important in human translation, and use all of them in N-gram counts (Papineni et al., 2002) or in measuring edit distances (Akiba et al, 2001; 2003). The variable part for different human translations is, in many cases, limited to a choice of a syntactic context for the stable core of content words. Therefore, more accurate MT evaluation results are obtained even with a single human reference, if the scores for matching N-grams (or

for edit distances) are adjusted with weights that reflect a word's significance in a given text.

The weighted N-gram model has been implemented as an MT evaluation toolkit (which includes a Perl script, example files and documentation) available at:

<http://www.comp.leeds.ac.uk/bogdan/ltv-mt-eval.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 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, one set of human translations was used to compute "significance" S-scores for each word in each of 100 texts, as suggested in (Babych, Hartley, Atwell, 2003). Section 3 describes the procedure of computing the S-scores.

In the second stage I 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 S-scores for every matched word.

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. 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.

## 3. Scores of statistical significance

The significance S-scores are based on the contrast of relative frequencies of words in a particular text and in the rest of the corpus of

human translations. They are computed using the following formula:

$$S_{word[txt]} = \ln \frac{(P_{word[txt]} - P_{word[rest-corp]}) \times N_{word[txts-not-found]}}{P_{word[all-corp]}}$$

where:

$S_{word[txt]}$  is the score of statistical significance for a particular word in a particular text;

$P_{word[txt]}$  is the relative frequency of the word in the text;

$P_{word[rest-corp]}$  is the relative frequency of the same word in the rest of the corpus, without this text;

$N_{word[txt-not-found]}$  is the proportion of texts in the corpus, where this word is not found (number of texts, where it is not found divided by number of texts in the corpus);

$P_{word[all-corp]}$  is the relative frequency of the word in the whole corpus, including this particular text. "Relative frequency" is (number of tokens of this word-type) / (total number of tokens).

The first factor ( $P_{word[txt]} - P_{word[rest-corp]}$ ) in this formula is the difference of relative frequencies in a particular text and in the rest of the corpus. Its value is very high for Named Entities (strings of proper nouns), which tend to recur in one text, but have a very low (often 0) frequency in the rest of the corpus. The higher the difference, the more significant is the word for this text.

The second factor  $N_{word[txt-not-found]}$  describes how evenly the word is distributed across the corpus: if it is concentrated in a small number of texts, the value is high and the word has more chances of becoming statistically significant for this particular text.

The third factor ( $1 / P_{word[all-corp]}$ ) normalises the words by their relative frequencies, so low-frequency and high-frequency words in language have an equal chance of becoming statistically significant. It is assumed that the author of a text has full "conceptual control" over words' significance.

The natural logarithm of the computed score is used to scale down the range of S-score values.

A threshold for distinguishing content words and function words was established by experiment:

$S\text{-score} = 1$ ;  $S\text{-score} < 1$  for (the majority of) functional words;  $S\text{-score} > 1$  for content words.

This threshold was found to distinguish content words and function words also in languages other

than English: it was tested for French and Russian, producing similar results.

Table 1 presents the results of ranking words in an example text according to their S-scores. A fragment of the text which produced this ranking is presented in Figure 1 below the table:

r	S	word	r	S	word
1	2,918	OPEC	8	1,844	total
1	2,918	Emirates	8	1,844	report
1	2,918	barrels	9	1,692	current
1	2,918	oil	10	1,593	price
1	2,918	quota	10	1,593	news
1	2,918	Subroto	11	1,470	recent
1	2,918	world	12	1,270	month
1	2,918	cartel	13	1,161	officials
1	2,918	war	14	0,972	because
1	2,918	ruler	15	0,805	million
1	2,918	petroleum	16	0,781	yesterday
1	2,918	markets	17	0,651	That
1	2,918	gestures	18	0,621	also
1	2,918	estimates	19	0,527	much
1	2,918	conciliatory	20	0,331	But
1	2,918	Zayed	21	0,291	over
1	2,918	UAE	22	0,007	from
1	2,918	Szabo	23	-0,079	there
1	2,918	Sheik	24	-0,126	after
1	2,918	Saudi	25	-0,233	their
1	2,918	Petroleum	26	-0,244	new
1	2,918	Dhabi	27	-0,284	had
1	2,918	Arabia	28	-0,411	as
1	2,918	Abu	29	-1,225	talks
2	2,719	output	30	-1,388	been
3	2,449	others	31	-1,594	at
3	2,449	manager	33	-1,844	on
3	2,449	government	34	-2,214	its
3	2,449	dropped	35	-3,411	for
3	2,449	declines	36	-3,707	with
3	2,449	agency	38	-4,238	The
4	2,375	day	39	-4,319	by
5	2,305	production	40	-4,458	Mr
6	2,096	well	41	-5,323	the
6	2,096	demand	42	-	a
7	1,880	concern	42	-	of

**Table 1. S-scores**

[...] After recent major price declines because of a sharp rise in OPEC production, total cartel output has now dropped as much as one million barrels a day -- from the year's peak of 20 million barrels a day early this month. And there are new indications that even the United Arab Emirates, OPEC's most notable quota violator, might soon cut its own output below a targeted 1.5 million barrels a day because of pressure from others in the cartel, including Saudi Arabia.

The Emirates aren't making any promises publicly. But its government did offer conciliatory gestures to OPEC yesterday. [...]

**Figure 1. Fragment of the text for ranking**

The assumption is that the ranking according to the S-score represents the “centrality” of given concepts for this text, which might be useful for a number of NLP applications, including automatic evaluation of MT.

The S-scores were generated for each word in each text in both sets of human translations available in the DARPA corpus.

#### 4. N-gram-based evaluation with S-score weights

In the second stage of the experiment the significance S-scores are used as weights for adjusting counts of matched N-grams when the output of a particular MT system is compared to a human reference translation.

The question that is specifically addressed is 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. But each time only a single reference translation was used. The scores for both runs were compared using a standard deviation measure.

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 N-gram \$ngr by 1:

```
$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 S-scores for words in every text in the corpus, similar to Table 1.

S-scores made a considerable contribution to the weights of all N-grams in the tested texts, reference texts, and to counts of N-grams found in both sets of texts.

Table 2 summarises this contribution for the matched N-grams. It gives an idea of how much the proposed approach relies on significance weights, and how much on the “heritage” of the baseline N-gram counts used by the BLEU method. The added weights to N-gram counts in tested and in reference translations represented about 97%–98% of the total score used for MT evaluation.

	<i>N-grams matched</i>	<i>Sum of S-weights</i>	<i>% added to N-gram no.</i>
<i>candied</i>	45074	1,654,396.0	97.2 %
<i>globalink</i>	41700	1,594,201.5	97.4 %
<i>ms</i>	44433	1,682,107.0	97.3 %
<i>reverse</i>	46403	1,762,911.3	97.4 %
<i>systran</i>	47102	1,799,162.3	97.4 %

**Table 2. Matched N-grams and S-scores**

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.

Table 3 summarises the following scores:

- 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. The first row in each case shows correlation *r*(2) (with 2 degrees of freedom) for the sets which include only scores

for MT systems, but not the “Expert” human translation; the second row shows correlation  $r(3)$  (with 3 degrees of freedom) when the scores for the human translation have been added to the set. 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
<i>HT-Expert</i> 0.921 / 0.852		0.7945	0.6685	0.7261
<i>CANDIDE</i> 0.677 / 0.455	0.3561	0.6996 0.7020	0.5067 0.5072	0.5877 0.5889
<i>GLOBALINK</i> 0.710 / 0.381	0.3199	0.6306 0.6318	0.4883 0.4876	0.5504 0.5504
<i>MS</i> 0.718 / 0.382	0.3003	0.6217 0.6201	0.5152 0.5111	0.5635 0.5603
<i>REVERSO</i> <i>NA / NA</i>	0.3823	0.6793 0.6805	0.5400 0.5389	0.6017 0.6015
<i>SYSTRAN</i> 0.789 / 0.508	0.4002	0.6850 0.6869	0.5511 0.5507	0.6108 0.6113
<i>Corr r(2) with [ade] – MT</i>	0.5918	0.0726 0.0708	<b>0.8347</b> <b>0.8271</b>	0.5686 0.5469
<i>Corr r(3) with [ade]MT&amp;HT</i>		0.8080	<b>0.9718</b>	0.9355
<i>Corr r(2) with [flu] – MT</i>	0.9807	0.8641 0.8618	0.8017 0.8440	<b>0.9802</b> <b>0.9894</b>
<i>Corr r(3)with [flu] MT&amp;HT</i>		0.9556	0.9819	<b>0.9965</b>

**Table 3. Evaluation scores**

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:  $r(2)=0.9807$ ,  $p > 0.05$ . 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:  $r(2)=0.8347$ ,  $p > 0.05$ ;  $r(2)=0.8271$ ,  $p > 0.05$ , but which becomes significant when the scores for the human translations are added to the set:  $r(3)=0.9718$ ,  $p < 0.01$ ).

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 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.

In the first place we assume that the reference translation is adequate to the original and we really estimate Adequacy of “monolingual translation” from “crummy” English into “normal” English. However, human subjects who evaluate Adequacy are bilingual and they use different sets of data for evaluating it: so whether or not the human reference translation is adequate to the original is questionable. Certainly, the baseline here is not a 1.0 score: note that the alternative human translation (which I use as reference in the second run of the experiment) scored not 1.0 but 0.921 and 0.852 on the Adequacy and Fluency scales respectively, according to human judges.

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:  $r(2)=0.9802$ ,  $p<0.05$ ;  $r(2)=0.9893$ ,  $p<0.01$ ;  $r(3)=0.9965$ ,  $p<0.01$ .

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, 2001:2) and 2 human reference translations, I 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, 2001:4) – 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.

Note that the correlation scores presented are highest if the evaluation unit is an entire corpus of translations produced by an MT system. For text-level evaluation, correlation is much lower. This may be due to the fact that human judges are not always consistent and when asked to score a text which contains fragments of variable quality, do this more or less randomly (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. This observation suggests that in order to get reliable scores at the text level one needs to let evaluators score at the level lower than text, i.e., on the level of individual segments, which on the whole may filter out the randomness of human intuitive judgments.

Automatic evaluation methods such as BLEU (Papineni et al., 2001), RED (Akiba et al., 2001), a method based on a parser performance on MT output (Rajman and Hartley, 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.

## 5. Stability of weighted evaluation scores

In previous sections I indicated that the weighted N-gram model improved the usability of

the MT evaluation tool, since only one human reference translation is required (apart from a monolingual corpus, which needs to be acquired only once and is assumed cheaper than producing multiple reference translations for each evaluated text).

The model also computes Recall and the Recall-based F-score, which have been found to straightforwardly correlate with different MT quality aspects. The inability to compute Recall in BLEU was the reason for a somewhat “one-sided” evaluation and generated unnecessary complications. The separation of Precision and Recall measures might also have an elegant linguistic interpretation, and develop a deeper understanding of the nature and links between different aspects of MT quality.

Central to both these issues is stability of scores across different runs of evaluation, when alternative human reference translations are used. In this section I compare stability of my results with stability of the baseline N-gram model with a single reference.

In order to carry out this comparison I 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. This comparison shows that stability of evaluation scores is improved by the use of significance weights.

In this stage of the experiment I 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 4. On the top of each “Sc-” (Score) cell is the result for the “Reference” human translation, on the top of the cell – the result for the “Expert” translation. Columns for Standard Deviations are grouped. The row “*Improved*” presents the percentage of change in stability for the weighted N-gram model.

	Systems	Sc-Baseline	StDev-basln	StDev-wtd	Sc-Weightd
P	candid	0.6364 0.6412	0.0034	0.0017	0.6996 0.7020
	globalink	0.5449 0.5469	0.0014	0.0008	0.6306 0.6318
	ms	0.5295 0.5287	0.0006	0.0011	0.6217 0.6201
	reverse	0.6030 0.6056	0.0018	0.0008	0.6793 0.6805
	systran	0.6072 0.6107	0.0025	0.0013	0.6850 0.6869
	<i>Ave SDev</i>		0.0019	0.0012	
	<i>Improved</i>				<b>+36.8%</b>
R	candid	0.6015 0.6000	0.0011	0.0004	0.5067 0.5072
	globalink	0.5564 0.5529	0.0025	0.0005	0.4883 0.4876
	ms	0.5929 0.5861	0.0048	0.0029	0.5152 0.5111
	reverse	0.6192 0.6157	0.0025	0.0008	0.5400 0.5389
	systran	0.6285 0.6259	0.0018	0.0003	0.5511 0.5507
	<i>Ave SDev</i>		0.0025	0.0010	
	<i>Improved</i>				<b>+60.0%</b>
F	candid	0.6184 0.6199	0.0011	0.0008	0.5877 0.5889
	globalink	0.5506 0.5499	0.0005	0	0.5504 0.5504
	ms	0.5594 0.5559	0.0025	0.0023	0.5635 0.5603
	reverse	0.6110 0.6106	0.0003	0.0001	0.6017 0.6015
	systran	0.6177 0.6182	0.0004	0.0004	0.6108 0.6113
	<i>Ave SDev</i>		0.0009	0.0007	
	<i>Improved</i>				<b>+22.2%</b>
	<i>All scores improved</i>				<b>+39.7%</b>

**Table 4. Stability of weighted N-gram scores.**

The baseline approach gives relatively stable results: the standard deviation was not greater than 0.005, which means that BLEU 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).

However, the Weighted N-gram model improved the stability of the baseline N-gram model even further: the standard deviation did not exceed 0.003, and the scores are about 40% more stable on average.

## 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 my 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.

Exploring applicability of this idea to different 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 2001 Bleu: a method for automatic evaluation of machine translation. IBM research report RC22176 (W0109-022) September 17, 2001
- Rajman, M. and T. Hartley. 2001. Automatically predicting MT systems ranking compatible with Fluency, Adequacy and Informativeness scores. *Proceedings of the 4<sup>th</sup> ISLE Workshop on MT Evaluation, MT Summit VIII*. Santiago de Compostela, September 2001. pp. 29-34.
- White, J., T. O’Connell and F. O’Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons and future approaches. *Proceedings of the 1<sup>st</sup> Conference of the Association for Machine Translation in the Americas*. Columbia, MD, October 1994. pp. 193-205.