Modelling legitimate translation variation for automatic evaluation of MT quality

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

Automatic methods for MT evaluation are often based on the assumption that MT quality is related to some kind of distance between the evaluated text and a professional human translation (e.g., an edit distance or the precision of matched N-grams). However, independently produced human translations are necessarily different, conveying the same content by dissimilar means. Such legitimate translation variation is a serious problem for distance-based evaluation methods, because mismatches do not necessarily mean degradation in MT quality. In this paper we explore the link between legitimate translation variation and statistical measures of a words salience within a given document, such as tf.idf scores. We show that the use of such scores extends the N-gram distance measures in a way that allows us to accurately predict multiple quality parameters of the text, such as translation adequacy and fluency. However legitimate translation variation also reveals fundamental limits on the applicability of distance-based MT evaluation methods and on data-driven architectures for MT.

Introduction

Automatic methods of Machine Translation evaluation aim at discovering formal metrics that correspond to intuitive human judgements about different aspects of MT quality, such as translation adequacy and fluency. Automatic evaluation tools enable MT developers and users to make quick judgements about MT quality without going through a lengthy and expensive process of human evaluation. Several automatic methods for MT evaluation have been proposed in recent years, e.g.: a method for predicting MT systems ranking by measuring performance of a parser on a target text (Rajman and Hartley 2001); the RED method, which measures edit distance between an evaluated text and a reference (Akiba et al., 2001); the BLEU method, which scores MT output by measuring a modified N-gram precision in relation to a set of human reference translations (Papineni et al. 2002).

Automatic MT evaluation methods that use human reference translations need to account for *legitimate translation variation* (LTV) – the fact that independent human translations of the same text often use different words and structures to convey the same content, which results in different sets of N-grams for such translations. In two independent human translations available in the DARPA-94 corpus the average overlap of unigrams in a text (about 350 words long) is approximately 72% for tokens and 68% for types. If N_{max} = 4, the overlap decreases to 46% for tokens and 44% for types.

In this paper we link LTV with the phenomenon of variable importance of translated units for MT evaluation. We put forward the suggestion that such differences in information load may be approximated by statistical weighting of words in reference translations with tf.idf scores (Salton and Lesk, 1968) or S-scores (Babych, Hartley, Atwell, 2003), which capture the "salience" of lexical items within a given document. We show that there is a noticeable difference in distribution of such significance weights for the words that are, respectively, variable or stable in independent human reference translations, although the relation between the significance weights and the LTV factor is not straightforward.

The proposed model yields extensions to the BLEU MT evaluation method, which have been implemented and tested on the DARPA-94 MT evaluation corpus of English translations produced by five MT systems and two human translators for 100 French news texts (White et al., 1994). The results show that for knowledge-based MT systems both methods produce similar scores compatible with human adequacy and fluency evaluation, but for a statistical MT system our method makes better a prediction of adequacy compared to the BLEU method. It was also established that for statistical MT output there is a greater gap between the precision and recall-based scores: human judgements about MT adequacy are in line with the latter, but not with the former. This reveals the importance of the recall-based measures (which are not generated by the BLEU method) for proper evaluation of data-driven MT architectures. We show that weighted recall-based scores are good indications of MT adequacy across different MT architectures.

Assumption of Reference Proximity

The key hypothesis behind methods that compute different kinds of distances between the human translations and MT output is the "assumption of reference proximity" (ARP), which states that "*the closer the machine translation is to a professional human translation, the better it is*" (Papineni et al., 2002: 311). Strictly speaking LTV undermines this assumption, since there is an ambiguity in interpreting deviations from the reference in the evaluated text. On the one hand these deviations may be the result of mistranslations, inadequate or nonsense translations or degraded fluency of MT. On the other hand they may be the result of choosing a legitimate alternative construction, which could be equally fluent, adequate and comprehensible.

A possible way of accounting for LTV is suggested by the BLEU method, which allows users to employ several human references. This reduces the ambiguity in interpreting deviations from the single human reference (there is a greater chance that a legitimate alternative will be found at least in one translation), but there is no guarantee that such set of references exhausts legitimate translation variants, or that deviations from an N-gram set

for all available references necessarily mean deterioration in MT quality. In addition this clearly increases the cost of MT evaluation, since multiple reference translations of the same text may be expensive to obtain.

Another disadvantage of using multiple human references is the so called "trouble with Recall" (Papineni et al., 2002:314): it only makes sense to compute *Precision* on a *union* of reference N-grams, because a good translation will use only one of the possible translation choices for a given unit, but not all of them. Still, Recall may contain important information about some aspects of MT quality. Intuitively this disadvantage means that, despite there being no proper translation equivalent for a certain concept which might be central for a given text, the MT system may still somehow "get away with it", without being directly punished for this omission.

Yet in practical terms, despite the above theoretical drawbacks, the ARP has been found to give good estimation of translation quality for mainstream knowledge-based commercial MT systems (even with a single reference). The relative number of legitimate and erroneous deviations from the reference appears to be relatively stable for MT systems built with the same architecture. If human translations produced by a native-speaker are included in the evaluation, the ARP approach still correctly ranks human translations higher than MT, although the difference in scores becomes much smaller.

Problems with ARP become more visible for "non-classic" types of texts, i.e., if we include data-driven MT systems, such as statistical MT, or non-native human translations into the evaluation set. In these cases the absence of a proper model for LTV cannot be compensated by other factors. With non-native human translation, a much greater proportion of mismatches "makes sense" and is judged useful by human evaluators. With statistical MT the situation is the opposite: relatively fewer mismatches actually "make sense" for human evaluators (and possibly the proportion of "spurious" matches is also relatively higher). Thus, when human evaluators compare the output of systems based on different architectures, the statistical MT system Candide is ranked higher with respect to its translation fluency than with respect to its adequacy (cf. the first column of Table 2). As a result present-day ARP-based methods consistently underestimate the usefulness of non-native human translation and overestimate the adequacy scores for statistical MT. Moreover, such "non-classic" texts cannot be judged using a single quality criterion. Different aspects of translation quality (such as adequacy and fluency) do not necessarily match up in the same translation. Therefore, a further challenge for ARP evaluation is the need to account for different quality criteria that may produce different rankings for evaluated systems.

The fact that statistical MT produces more fluent, but still not highly adequate translation indicates the need for ARP-based evaluation tools to account for different aspects of MT quality. An ARP-based model should predict which terms are more important for evaluation and which terms might be subject to greater LTV.

LTV and frequency weighting scores

To account for LTV within ARP-based MT evaluation models we propose to extend proximity scores with statistical weights of term "salience" or "significance" within a text, such as tf.idf and S-scores. This extension is based on the assumption that these measures approximate the relative importance of lexical items for human translators and evaluators, and this will be necessarily reflected in LTV across different human translations. In our experiment we computed the tf.idf and S-scores for each lexical type in each text in the DARPA corpus as follows:

The tf.idf scores:

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

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

The tf.idf and S-scores were computed on the basis of both reference translations.

To establish the impact of these significance scores on LTV, we divided the unigrams from the two human translations into three classes: those found in both translations (the intersection set of unigrams) and those found in only one of the translations (two differences sets of unigrams). We examined the distribution of tokens with different significance scores in each of these classes. For the intersection class we did two calculations on the basis of each of the human translations. For the difference class we did the calculation only on the basis of the "native" reference translation.

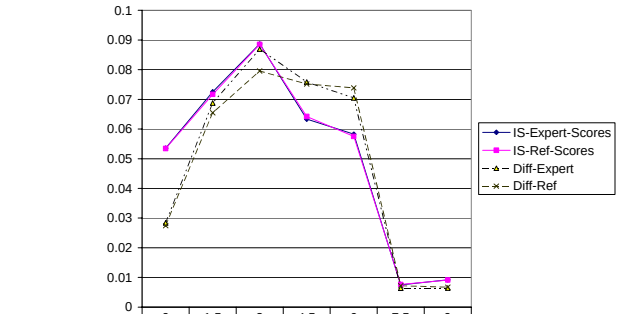
Since the intersection set (IS) of unigrams is larger than the difference sets, we compared the average tf.idf and S-scores and frequency polygons of scores normalised by the size of each set. Table 1 presents the average scores for the sets:

	tf.idf score	S-score
IS-Expert-Scores	2.6057	1.9825
IS-Ref-Scores	2.6146	2.0011
Diff-Expert	2.8290	2.2206
Diff-Ref	2.9200	2.3046

Table 1. Average scores: Intersection and Difference sets

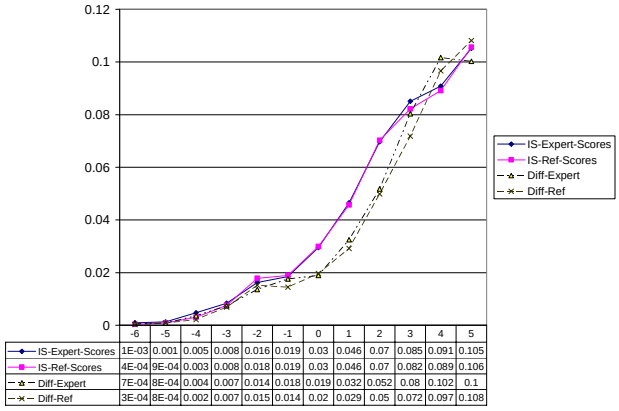
These results are surprising because terms in the difference sets (those which were found to undergo LTV) have somewhat higher average significance scores that the supposedly more stable terms in the intersection sets. (Intuitively one may be inclined to believe that more significant words, such as content words, should be also

more stable, and translation variation may be mostly due to the choice of low-salience functional words or different morpho-syntactic perspectives for a sentence).
This means that stable words across human translations are somewhat less "salient" than the words with variable translation equivalents. Frequency polygons for each of these scores describe the distribution of significance scores for the intersection and difference N-gram sets. Figures 3 and 4 compare the frequency polygons (normalised by the size of each N-gram set) for each set weighted by tf.idf or by S-scores.



	0	1.5	3	4.5	6	7.5	9
IS-Expert-Scores	0.0536	0.0725	0.0887	0.0633	0.0583	0.0074	0.0092
IS-Ref-Scores	0.0534	0.0717	0.0885	0.0642	0.0575	0.0076	0.0091
Diff-Expert	0.0284	0.0688	0.0869	0.0758	0.0705	0.0064	0.0064
Diff-Ref	0.0274	0.0655	0.0795	0.0752	0.0739	0.0072	0.0067

Figure 3. Frequency polygons weighted by tf.idf



	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5
IS-Expert-Scores	1E-03	0.001	0.005	0.008	0.016	0.019	0.03	0.046	0.07	0.085	0.091	0.105
IS-Ref-Scores	4E-04	9E-04	0.003	0.008	0.018	0.019	0.03	0.046	0.07	0.082	0.089	0.106
Diff-Expert	7E-04	8E-04	0.004	0.007	0.014	0.018	0.019	0.032	0.052	0.08	0.102	0.1
Diff-Ref	3E-04	8E-04	0.002	0.007	0.015	0.014	0.02	0.029	0.05	0.072	0.097	0.108

Figure 4. Frequency polygons weighted by the S-score

It can be seen from the charts that terms with different salience scores vary in their stability across independent human translations. In the first place there is no significant difference for "under-represented" words (S-score < 0). The words with low and average salience scores (0 < tf.idf < 3; 0 < S-score <3) constitute the majority of the words used in texts. They tend to be much more frequent in the intersection set, i.e., they are more stable across independent human translations. On the contrary, a relatively small number of highly salient words (S-score > 3; if.idf > 3) become more frequent in the difference set, therefore being subject to the greater translation variation.

The threshold of S-score = 1 accurately distinguishes function words from content words, and it can be seen that the majority of function words show clear stability, which is not substantially different from the stability of content words that are highly frequent in a corpus.

These results suggest that the words which are not salient within a given text usually have some optimal translation equivalents. Different human translators usually agree on these equivalents.

However, individual human translators are also very consistent in using words which are subject to great translation variation, which makes these words statistically salient. This means that highly significant units typically do not have ready translation solutions and require some "artistic creativity" on the part of human translators. Such words also give the translators a degree of freedom, making translation to some extent a creative process, even supposedly "non-computable" or "non-algorithmic" (cf. Penrose, 1990), which involves creative invention of translation strategies.

Finding out a proper translation strategy for such unstable words is very important for the general quality of the text, since highly salient words make the biggest contribution to the texts general content, and matter most of all in evaluation of the text quality by human judges[1].

The results presented suggest that there is a potentially serious problem for ARP-based approaches to MT evaluation: the most important terms in translation are the most unstable ones, which may not be necessarily present in any number of human reference translations. However, this problem may be partly solved by assigning different weights to highly salient and low significant N-gram matches in a reference translation and the evaluated text.

Using significance weights for MT evaluation

The described properties of significance weights suggest that they could adjust distance measures between the evaluated text and the reference in ARP-based evaluation models. For the N-gram distances this means that matches of more salient concepts matter more for human evaluators, so tf.idf and S-score could act as weights for the counts of matches. The model of weighted N-grams distinguishes different possible "angles" of matches between the evaluated text and the human reference, which can be visualised by the two diagrams in Figure 5. On the diagrams the set of reference N-grams is divided into the following subsets:

– *dh* – N-grams in the difference set that have high salience scores ("different-high" N-grams);
– *dl* – N-grams in the difference set that have low salience scores ("different low" N-grams);
– *cl* – N-grams in the intersection set with low salience scores ("common low" N-grams);

1 The following sentence is an example of LTV related to the absence of a clear-cut translation strategy; transformations are applied differently by the two human translators:

ORI : Le président de la chambre d'accusation doit ***rendre*** un ***avis*** de clôture, ouvrant un ***délai*** de vingt jours pour les ***requêtes*** des diverses parties, suivi d'un arrêt de "soit communiqué" pour le ***règlement*** du dossier par la parquet général de Lyon.

REF : The Director of the Public Prosecutor's Office must ***give*** a closing ***decision***, which will open a 20-day ***period*** for the various parties to file ***petitions***, after which no papers may be sent to the public prosecutor so that the Office of the Public Prosecutor of Lyon can ***prepare*** the case.
EXP : The presiding judge of the Court of Criminal Appeals is to ***render*** a closing ***opinion***, thus establishing a twenty-day ***deadline*** for ***requests*** from the various parties, followed by a "may it be communicated" order for ***settlement*** of the case by the Lyon public prosecutor's office.

– *ch* – N-grams in the intersection set with high salience scores ("common high" N-grams);

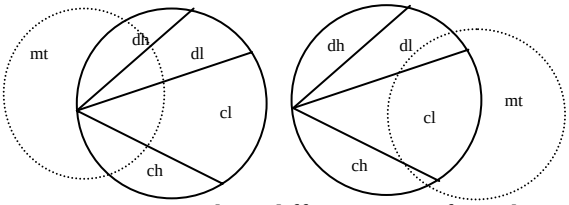


Figure 5. Distinguishing different types of matches

N-grams from the MT output intersect with the reference set at different angles. For instance, even if total count of the matched N-grams is the same, the output of knowledge-based MT systems probably intersecst with more important, i.e., higher scored N-grams, while the output of statistical MT intersects with the set of reference N-grams "from the other side", where mostly low score N-grams are matched.

The proposed LTV model provides a framework for fine-tuning MT evaluation tools with respect to different quality criteria, allowing the user to modify the range and magnitude of the significance scores involved in MT evaluation.

For the four MT systems available in the DARPA corpus we experimentally established that the highest correlation with human evaluation scores is achieved with the following settings:

– for fluency the precision of N-grams weighted by tf.idf scores gives the best results (Piersons correlation coefficient *r* is around 0.99 for both human translations used as a reference)
– for adequacy the recall of N-grams weighted by S-scores gives the best results (*r* is around 0.90 for both human reference translations).

Table 2 shows the evaluation results with the described settings for these two quality parameters using a single human reference translation each time and N-gram size = 4. The table also compares the weighted evaluation scores and the baseline scores generated by the BLEU method.

System [ade] / [flu]	**BLEU [1&2]**	**Prec.+tf.idf (w) 1/2**	**Recall+S-sc. (w) 1/2**
CANDIDE 0.677 / 0.455	0.3561	0.5242 0.5176	0.2553 0.2554
GLOBALINK 0.710 / 0.381	0.3199	0.4905 0.4890	0.2464 0.2493
MS 0.718 / 0.382	0.3003	0.4919 0.4902	0.2635 0.2679
SYSTRAN 0.789 / 0.508	0.4002	0.5442 0.5375	0.3034 0.3022
*Corr **r(2)** with **[ade]** – MT*	0.5918	0.5248 0.5561	**0.9069** **0.9215**
*Corr **r(2)** with **[flu]** – MT*	**0.9807**	**0.9987** **0.9998**	0.8022 0.7499

Table 2. LTV-aware MT evaluation results

The described approach has been implemented in an MT evaluation toolkit which includes documentation, a Perl script and frequency lists of words from DARPA-94 corpus that are used for computing tf.idf and S-scores for terms in the evaluated text. Users have an option of supplying their own corpus or frequency lists in the specified format. The toolkit is available at the URL:

http://www.comp.leeds.ac.uk/bogdan/ltv-mt-eval.html

Conclusions and future work

The discovered difference in the tf.idf and S-scores for terms that are subject to various degrees of translation variation indicates that there is a link between the potential stability of units across independent human translations and their "salience" within a given text. Highly significant words, which are consistently used within a single translation, were found to be the most unstable across different translations. The possible reason for this fact could be that translation of significant units typically requires invention of some novel translation strategy.

The results also indicate that there exist fundamental limits on using data-driven approaches to MT, since the proper translation for the most important units in text may be not present in the corpus of available translations. Discovering the necessary translation equivalent might involve a degree of inventiveness and genuine intelligence.

Future work will involve testing the applicability of our method for highly-inflected languages, where N-gram scarcity is higher, finding a linguistic interpretation of the significance weights, and establishing the potential limits of legitimate variation across multiple human translations of a single text.

References

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

Babych, B; Hartley, A.; Atwell, E. 2003. Statistical Model-ling of MT output corpora for Information Extrac-tion. In: Proceedings of the Corpus Linguistics 2003 conference, edited by Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery. Lan-caster 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 ma-chine translation. Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

Penrose, R. 1989. The Emperors New Mind. Oxford University Press.

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.

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. OConnell and F. OMara. 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.