Training a Super Model Look-Alike: Featuring Edit Distance, N-Gram Occurrence, and One Reference Translation

Eva Forsbom

Department of Linguistics, Uppsala University
Uppsala, Sweden
evafo@stp.ling.uu.se

Abstract

Two string comparison measures, edit distance and n-gram co-occurrence, are tested for automatic evaluation of translation quality, where the quality is compared to one or several reference translations. The measures are tested in combination for diagnostic evaluation on segments. Both measures have been used for evaluation of translation quality before, but for another evaluation purpose (performance) and with another granularity (system). Preliminary experiments showed that the measures are not portable without redefinitions, so two new measures are defined, WAFT and NEVA. The new measures could be applied for both purposes and granularities.

1 Motivation

When translating texts for external publication, an important object of evaluation, for manual as well as machine translation, is translation quality, such as adequacy and fluency, particularly if the readers have to perform a task or base their decisions on information in the publication, e.g. automotive service literature.

So far, quality evaluation has mostly been done by human evaluators, who can actually understand the text and its translated form, which computers cannot. But skilled human evaluators cost a lot, and are not always readily available, so it seems a good idea to save their efforts to one or possibly a few formal evaluation rounds when the system has been trained for a while. For the training periods, however, we need some way of approximating their judgments.

Evaluating translation quality often involves comparisons of the translated text against the source text (adequacy) or against other translations of the same text (adequacy and fluency). Assuming we have a training corpus of source texts aligned with reference translations, quality evaluation could be done by string comparison.

A simple way of doing string comparison is to count the edit distance between the source and target texts, i.e. the minimum number of edit operations it takes to turn the first into the other. Such a comparison has been used in natural language applications for a long time, and in evaluations of such applications (see Section 2.1).

In recent machine translation evaluation forums, e.g. those performed by DARPA, it has also been shown that n-gram measures correspond closely to human evaluations of adequacy and fluency for machine translations, at least for ranking systems.

These evaluations have mainly been made for fully-trained systems, on news texts, with English as the target language, and with several reference translations to evaluate against. In the ISLE taxonomy (ISLE, 2002), this would correspond to a declarative evaluation.

For most systems under development, however, reference translations are scarce—there is often only one reference translation available—and if a system should be trained on a specific domain and text type, it is not so relevant to reuse general test sets for evaluation, should they exist for the language pair in question, since the object of evaluation during training is to see how well the system performs on the particular domain and text type, i.e. (1) whether a change in the system makes the translation of the training text better or worse, (2) how much the translations of the system versions differ, and (3) what the difference is. In the ISLE taxonomy, the first two would correspond to progressive (internal) evaluation, and the third to diagnostic evaluation.

In this paper, we will focus on the applicability of the edit distance and n-gram occurrence measures for another text type than news texts (Section 3), for evaluations with only one reference translation (Section 4), and for diagnostic evaluations (Section 5).

2 Background

Evaluations of translation quality take time, even if they are done partly automatically, both with regard to the actual processing and with regard to learning time for the user. Therefore, it would be ideal to find measures that are applicable for various evaluation purposes, e.g. declarative, progressive and diagnostic evaluation, and with various granularity. Prior evaluations using edit distance and n-gram measures have mainly concentrated on ranking systems or system versions at the system or document level, i.e. the systems' performance on a complete corpus and on the individual documents in the corpus, respectively. We would like to see if these measures are equally applicable for grading (or scoring) a system at the segment level, i.e. the individual text segments (sentences, headlines, etc.) in the corpus. The measures used for this are described below.

2.1 Edit Distance

Word accuracy (WA) has recently been introduced in evaluations of MT systems (Alshawi et al., 1998), and seem to correspond well with human evaluations. The idea behind the measure is that it would approximate a post-editor, in that it is based on edit distance, i.e. the minimum number of deletions, substitutions and insertions of words needed to turn the candidate translation into the reference translation, as defined in equation 1.

$$\mathrm{WA} = \left(1 - \frac{d+s+i}{r}\right) \qquad (1)$$

where

$$\begin{aligned} d &= \text{deletions} \\ s &= \text{substitutions} \\ i &= \text{insertions} \\ r &= \text{length of reference} \end{aligned}$$

For this test, we used `Scoring Toolkit`[1] from NIST. In the case of several reference translations, the one producing the best score is used.

2.2 N-Gram Co-Occurrence

Lately, n-gram co-occurrence measures such as NIST (Doddington, 2002) and BLEU (Papineni et al., 2001) have been used in machine translation evaluations, and they seem to correlate well with human judgments on adequacy and fluency. The idea behind these n-gram measures is that 1-grams occurring in both the candidate and reference translation(s) reflect accurate terminology, while the higher n-grams reflect fluency.

For our study, we chose BLEU, for although the NIST measure is good for ranking, it is less so for grading, since it has no specific upper boundary to tell if the candidate translation is identical to one of the reference translations. Furthermore, as the NIST measure uses information weights, it is not certain that two candidate segments of equal length and identical to their respective reference translation will get the same score. The candidate *Number*, for example, with the reference *Number*, receives the score 4.6267, while the candidate *Designation*, with the reference *Designation*, receives the score 8.0311, using the `mteval-kit`[2] from NIST (the same program as we used for computing BLEU).

BLEU, on the other hand, is bounded by 0 and 1, where 0 means no similarity between the candidate translation and the reference translation(s), and 1 means full similarity. The measure, defined in equation 2, counts the number of n-grams of the candidate translation which occur in the reference translation(s) and gives them equal weight. If the candidate is shorter than the reference(s), there is a brevity penalty.

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad (2)$$

where

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases}$$

$$\begin{aligned} r &= \text{length of reference} \\ c &= \text{length of candidate} \end{aligned}$$

$$\begin{aligned} N &= 4 \\ w &= \tfrac{1}{N} \\ p &= \frac{\sum_{C \in \{Cand\}} \sum_{n-gram \in \{C\}} Count_{clip}(n)}{\sum_{C \in \{Cand\}} \sum_{n-gram \in \{C\}} Count(n)} \end{aligned}$$

[1] `http://www.nist.gov/speech/tools/`

[2] `http://www.nist.gov/speech/tests/mt/`

2.3 Corpora for Testing

To test the applicability of these measures at the segment level, we have used two test corpora: a subset of the test corpus for the LREC'02 MTEval workshop[3] and a subset of the MATS[4] corpus.

The LREC set consists of the machine-translated part of the workshop corpus, i.e. candidate translations from 7 MT-systems (or versions of MT-systems) in four similarity sets: 7; 8, 9; 10, 11, 13; 12 (Popescu-Belis, 2002), where each set groups candidates from all versions of a particular system. The candidates are translation from French to English of two news articles having 37 segments in total. Each article has 1 reference translation, and in addition, one article has 6 and the other one has 4 manual translations to evaluate against (hereafter referred to as "6/4 references").

The MATS set consists of the fully-translated segments from 14 documents of the training part of the MATS corpus, translated from Swedish to English by the MATS system (Sågvall Hein et al., 2002), all in all 1736 segments (or 943 unique segments).

3 Applicability for another Text Type

MT evaluations using edit distance and n-gram measures have mainly been made on news texts, which are of a quite different text type than automotive service literature. Technical manuals in general, tend to have many short segments (e.g. list items and table cells), have many compounds (e.g. noun clusters), and be more repetitive.

Preliminary tests on the MATS set revealed some irregularities in both WA and BLEU at the segment level, particularly for short segments. Thus, the measures needed redefinition in order to be useful for diagnostic evaluation on this text type.

3.1 Edit Distance Redefinition

The problem with the WA measure defined in equation 1 is that the length of the reference translation is used as denominator. But, if that is shorter than the candidate, it could result in a word accuracy score less than 0, as in the following example from the MATS set (WA= -1):

Src:	Tätningsring
Cand:	Sealing ring
Ref:	Seal

In other applications where WA has been used, such as in Automatic Speech Recognition, the candidate and reference are probably of the same length, so there is no need to account for differences in length. In translation, however, candidate and reference translations often differ in length. When there is a difference, there will always be a corresponding number of insertions or deletions to account for the difference, so it would be better to use the longest of the candidate and reference translation as the denominator. That way, the value will always be between 0 and 1. We therefore used the revised word accuracy measure, Word Accuracy For Translation (WAFT), described in equation 3.

$$\mathrm{WAFT} = \left(1 - \frac{d+s+i}{\max(r,c)}\right) \tag{3}$$

The LREC set was used in order to verify in some respect that the changes are not only applicable to the MATS corpus. (For a full verification we would, of course, need a larger corpus.) The resulting WAFT scores for the LREC set evaluated against all 6/4 reference translations at system and document level are shown in Figure 1. They are slightly higher than the ones for WA (see Figure 2), but the ranking and similarity sets are basically the same, except for systems 8 and 9 on document 1, which were the only ones producing negative scores for WA in this set. The measures also correlate well at the segment level, except for cases where WA is negative or close to zero (see Figure 3), which were the ones suffering from the length difference, so the new measure seem to have remedied the problem without causing another.

3.2 N-Gram Co-Occurrence Redefinition

The problems with the BLEU measure described in equation 2 at the segment level are twofold. Firstly, it is not defined for candidate segments shorter than 4 words (N=4), since the denominator of p_n would be 0 for the 4-grams (and 3-grams and 2-grams, depending on the number of words in the segment). This means that segments such as the following example from the MATS corpus are not handled cor-

[3] Language Resources and Evaluation 2002: Workshop on Machine Translation Evaluation: Human Evaluators Meet Automated Metrics, `http://issco-www.unige.ch/projects/isle/mteval-may02/`

[4] Methodology and Application of a Translation System, `http://stp.ling.uu.se/mats/`

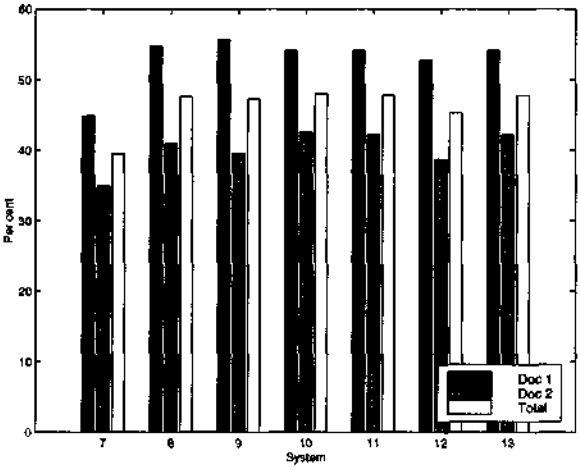

Figure 1: LREC set for 6/4 references using WAFT.

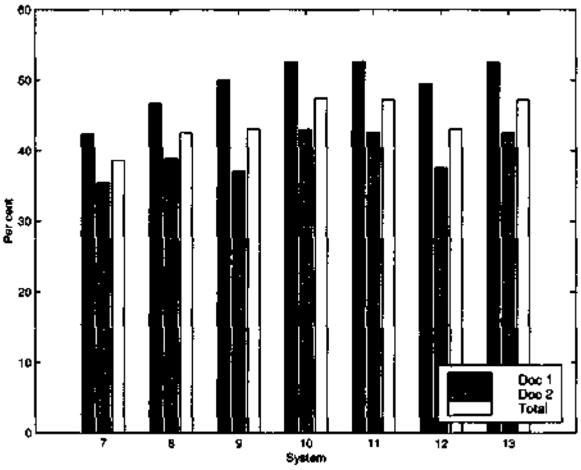

Figure 2: LREC set for 6/4 references using WA.

rectly[5]:

Src: Antal
Cand: Number
Ref: Number

Secondly, it is not defined for candidate segments not containing any 4-grams (or 3-grams or 2-grams), since p_n would be 0 and log 0 is not defined. Segments such as the following example from the MATS corpus are not handled correctly[6]:

Src: Ledningsnät för bränslepump
Cand: Cable harness for fuel pump
Ref: Fuel pump cable harness

At the document and system level, these problems rarely occur (if ever), so the measure still works for evaluations at those levels. At the segment level, however, and for technical manuals, in particular, they occur frequently. In the revised n-gram co-occurrence measure, N-gram EVAluation (NEVA), defined in equation 4, the problem of short segments is addressed by the redefinition of N, which means that the counting is only done for 4-grams (assuming N_{max} is 4) if the candidate segment length is 4 or more, and for the n-grams occurring in the segment if it is shorter.

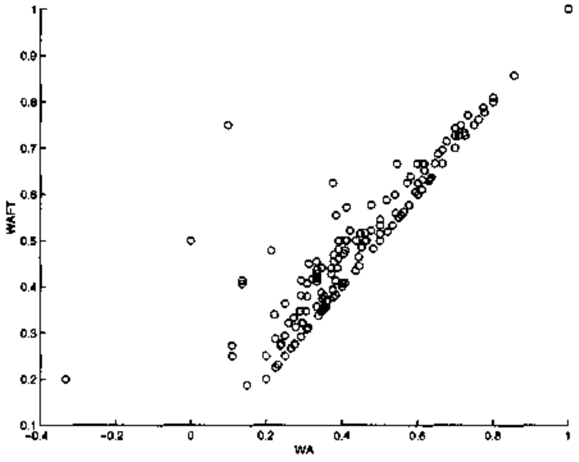

Figure 3: LREC set segment correlations between WA and WAFT for 6/4 reference translations.

The problem of non-occurring 4-, 3-, or 2-grams is handled in NEVA by leaving out the exp and log functions of BLEU. This is a simplification of the measure, based on the assumption that since exp(log x) is x, the new score would still be in the same order of magnitude and retain its relevant ingredients.

$$\text{NEVA} = \text{BP} \cdot \sum_{n=1}^{N} w_n p_n \tag{4}$$

where

$$\text{N} = \begin{cases} N_{max} & \text{if } c \geq N_{max} \\ c & \text{if } c < N_{max} \end{cases}$$

The LREC set was used to verify that the changes are not only applicable to the MATS corpus. The resulting NEVA scores for the LREC set evaluated against all 6/4 reference translations at system and document level are shown in Figure 4. They are slightly higher than the ones for BLEU (see Figure 5), but the ranking and similarity sets are basically the same. The measures also correlate well at

[5] `mteval` reports such cases as 0.

[6] `mteval` reports such cases as 0.

the segment level, except for cases where BLEU is zero (see Figure 6), which were the ones suffering from the two problems, so the new measure seems to have remedied the problem without causing another.

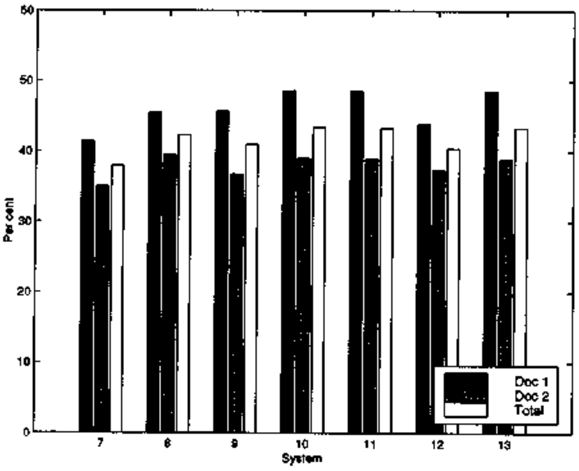

Figure 4: LREC set for 6/4 references using NEVA.

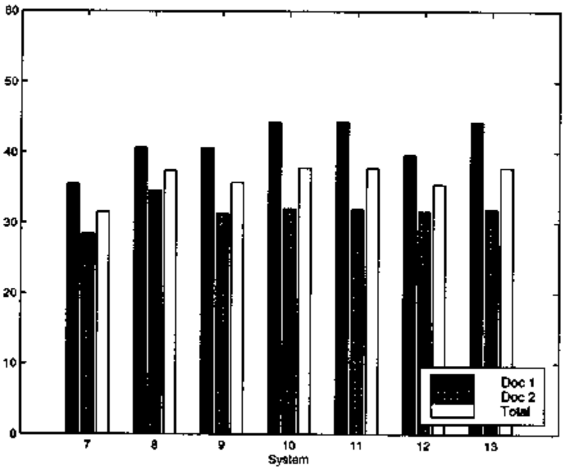

Figure 5: LREC set for 6/4 references using BLEU.

3.3 Repetitiveness Restriction

Another issue to be taken into consideration for technical manuals is how to deal with repetitiveness. Manuals, in contrast to news articles, often have many duplicate segments. And, while a human may translate identical source segments in different ways, a machine translation system will not, so if we would include all occurrences of such identically translated segments in the evaluation, the system could get overly favoured or penalised, depending on the score and number of occurrences of the particular segment. It would therefore be better to only include unique segments in a quality evaluation. By unique, we mean those that are translated in exactly the same way in the reference translation, not those that are translated differently (intendedly or not).

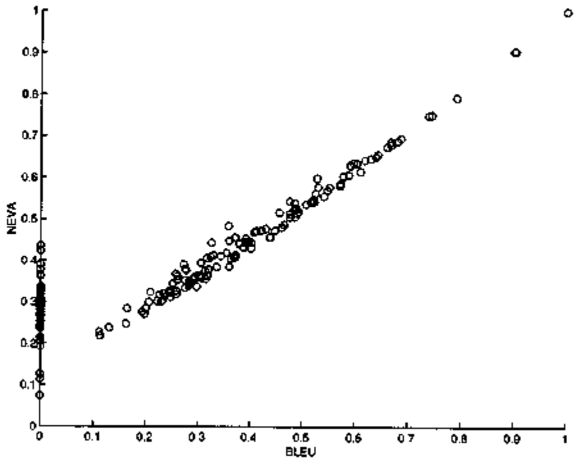

Figure 6: LREC set segment correlations between BLEU and NEVA for 6/4 references.

If we would like to still be able to evaluate at the document level, however, we have to keep one occurrence in every document in which the duplicate occurs, i.e. document-unique segments. Scoring level differences between evaluations including all segments and including only unique segments would probably be rather small (cf. Figure 7 for WAFT), since segments are concatenated for evaluations at the document and system levels for both measures.

4 Applicability for One Reference Translation

Unfortunately, there are few cases where several reference translations are available; in most cases there is only a single reference to be had. So, do the findings above concerning several reference translations also apply to a single reference? We would expect the grading (or scores) to be lower, since the probability of finding a (partially) matching reference segment is lower. We would also expect ranking of systems or system versions to be similar at the higher levels, if the single reference translation is comparable in quality to the other references, since the higher

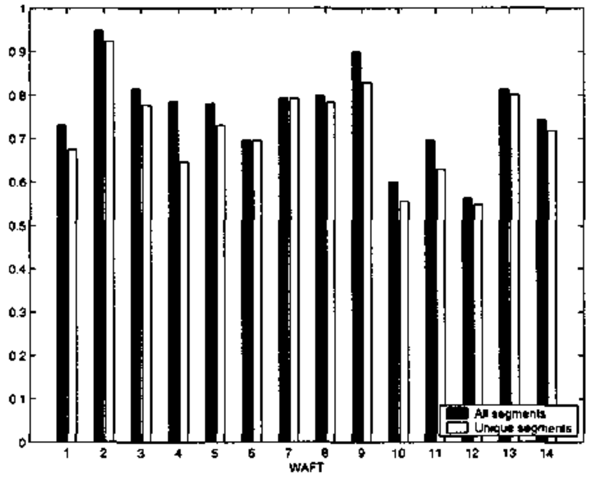

Figure 7: MATS set document scores for WAFT (all vs. unique segments).

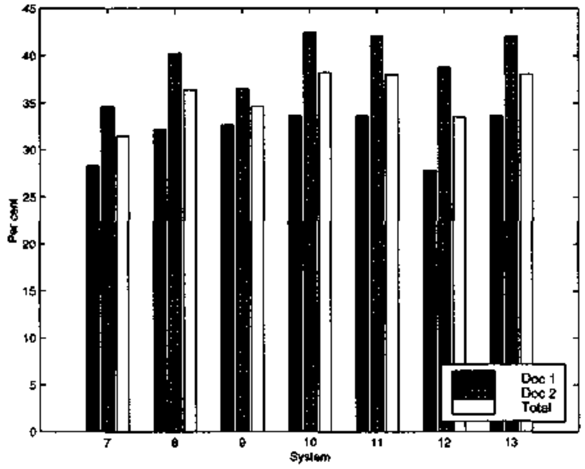

Figure 8: LREC set for 1 reference using WAFT.

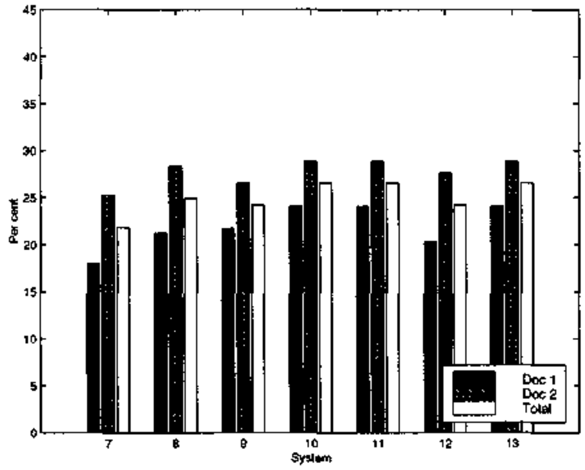

Figure 9: LREC set for 1 reference using NEVA.

levels measure the general translation quality of a system compared to the general quality of the reference translation.

The LREC set was used to test the applicability of the edit distance and n-gram occurrence measures for only one reference translation. Both the WAFT (Figure 8) and NEVA (Figure 9) scores for the LREC set evaluated against the single reference translation at system and document level were lower than the ones for 6/4 references (Figures 1 and 4). This was particularly true for NEVA. The rankings and similarity sets were basically the same, except for systems 8 and 9, while the scoring level for document 1 and 2 were reversed. This could either be because the reference translation for document 2 is of better quality than the one for document 1, or have something to do with the fact that document 2 has much longer segments. NEVA correlated better at all levels: system, document and segment (see Table 1), which could mean that NEVA is less sensitive to the quality of the reference than WAFT, but this has to be investigated further.

4.1 Creating the Super Model

When using one single reference translation, it is particularly important to consider the translation quality of the reference translation, since both scoring level and ranking are affected by it.

When a system is trained on a new domain or text type for external publication, previous translations for publishing are often the sole reference translation available on which to model the translation. Although published material generally are of good, or at least acceptable, quality, it is seldom free of faults, particularly not if produced within time and money constraints. Spelling errors, inconsistent use of terms and grammatical errors can occur, and we would not like to train the system on them. Another problem is that good translations are not always the best translations to train the system on, for example, if the translations contain explanations not found in the source texts, and such world knowledge is beyond the scope of the system.

Instead of asking translators to create other versions, we could, after training the system a while, ask them as evaluators to accept or reject system

Level	WAFT	NEVA
System	0.8589	0.9857
Document 1	0.6854	0.9983
Document 2	0.9348	0.9632
Segment	0.6215	0.7274

Table 1: LREC set correlations between WAFT scores for 1 and 6/4 references.

versions of segments where the current reference translation displays problems of the kind described above. Accepted versions might then be added in "cloned" versions of the original reference, to create a super model, much in the spirit of the Eval - Trans (Nießen et al., 2000) tool, but using only fully-accepted segments.

For the MATS corpus, there is only one reference translation, i.e. the one used for external publication. The source documents are subject to a controlled language, Scania Swedish (Almqvist and Sågvall Hein, 2000), while the target documents are not, as yet, although work is under progress (Sågvall Hein et al., 2002). This means that there are some inconsistencies left in the reference translation used. *Felsökning,* for example, is translated both as *Troubleshooting* and *Trouble shooting* in the reference, while the system always produces the former, i.e. the preferred form.

The advantage with the super model method is that the original reference is kept intact, and could still be used for formal evaluations where more than one system are involved, possibly in conjunction with clones of inconsistency corrections which are agreed upon.

5 Applicability for Diagnostic Evaluation

Both edit distance and n-gram co-occurrence measures are based on string comparison, and measures basically the same thing, but their characteristics differ somewhat, a difference which could possibly be used for diagnostic evaluation at the segment level. Compared to NEVA, WAFT generally gives a higher score, for example. But there are other differences, too.

The main weakness of edit distance measures is that they are very sensitive to reversed word order. The following candidate translation with its corresponding reference translation, for example, will get a WAFT score of 0, since the edit distance is as long as the segment length.

Src: Cylinder, underdel
Cand: Bottom cylinder
Ref: Cylinder bottom

N-gram co-occurrence measures, on the other hand, are very sensitive to word level errors, particularly if the word in question is located mid-segment and thus should partake in all possible n-grams. The following candidate translation with its corresponding reference translation will get a NEVA score of 0.3250, since the mismatched word *check*[7] breaks up the possible 4- and 3-grams:

Src: Kontrollera backventilen.
Cand: Check the check valve.
Ref: Check the non-return valve.

Mismatched words at the ends of a segment are not penalised as hard, as in the following example[8] (NEVA=0.4792):

Src: Generator och remspännare
Cand: Alternator and belt tensioners
Ref: Alternator and belt tensioner

As has been mentioned earlier, NEVA showed a much lower score when used with only one reference translation for the LREC set (see Section 4), which could be a consequence of its sensitivity to word errors.

Using knowledge of the measures' weak points, scoring levels and differences of length between the candidate and reference translation, it would be possible to single out segments with certain error types in a diagnostic evaluation. In the MATS set, for example, all segments where the NEVA scores were greater than the WAFT scores displayed a reversed word order problem, as in the following example, where NEVA is 0.3250 and WAFT is 0:

Src: Magnetventiler för insprutningstidpunkt
Cand: Solenoid valves for injection timing
Ref: Injection timing solenoid valves

When computing the edit distance using the dynamic programming technique, it is possible to backtrack the edit distance computation and create an edit operation alignment table (Navarro, 2001), which can be used for finding confusion pairs (or substitutions) such as variant forms and synonyms *(clip/clamp),* inflectional errors *(tensioner/tensioners),* or word errors *(in/into).*

Inserted and deleted words could point out differ-

[7] *Check valve* is the preferred variant.

[8] *Remspännare* is ambiguous in number.

ences in definiteness *(the),* or word order changes, which could be used in conjunction with the WAFT-NEVA difference to find out what parts were changed, e.g. *of* or *for* for splitted noun clusters. Insertions and deletions could also point to specifications and generalisations, if the inserted or deleted word corresponds to a nominal modifier.

6 Conclusions

In this paper, we focused on the applicability of edit distance and n-gram co-occurrence measures for evaluations of translation quality, in particular for technical manuals, one reference translation, and diagnostic evaluation.

We found that the measures were applicable for those purposes, although currently used measures for edit distance (WA) and n-gram occurrence (BLEU) needed to be redefined in order to handle evaluations at the segment level as well as they do at the document and system level, and to handle the characteristics of technical manuals. The redefined measures, WAFT and NEVA, respectively, gave slightly higher scores than WA and BLEU, but did not alter the ranking.

We also found that although the measures have a higher scoring level when used together with several reference translations, they are still able to rank at all levels when used together with only one reference translation. NEVA had higher correlation values than WAFT, which seems to suggest that NEVA is less sensitive to the number of references used.

Although both measures are based on string comparison, they differ in their sensitivity to certain error types: WAFT is more sensitive to word order differences, while NEVA is more sensitive to word-level errors. WAFT also has a higher scoring level in general. These differences could be used to single out certain error types in a diagnostic evaluation or in correcting inconsistencies in a single reference translation to make it more appropriate for its purpose. Further testing needs to be done on how the measures can help in diagnostic evaluation.

Acknowledgements

The project was supported in part by VINNOVA (Swedish Agency for Innovation Systems), contract no. 341-2001-04917.

References

I. Almqvist and A. Sågvall Hein. 2000. A language checker of controlled language and its integration in a documentation and translation workflow. In *Proceedings of the 22nd International Conference on Translating and the Computer,* Translating and the Computer 22, Aslib/IMI, London.

H. Alshawi, S. Bangalore, and S. Douglas. 1998. Automatic acquisition of hierarchical transduction models for machine translation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL'98),* pp. 41-47, Montreal, Canada.

G. Doddington. 2002. Automatic evaluation of language translation using n-gram co-occurrence statistics, May 27. Talk presented at Workshop on MT Evaluation: Human Evaluators Meet Automated Metrics (LREC'02).

ISLE. 2002. Taxonomy for MT evaluation. [http://www.issco.unige.ch/projects/isle/taxonomy2]

G. Navarro. 2001. A guided tour to approximate string matching. *ACM Computing Surveys,* 33(1):31-88.

S. Nießen, F.J. Och, G. Leusch, and H. Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC'00),* pp. 39-45, Athens, Greece.

K. Papineni, S. Roukos, T Ward, and W.-J. Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. IBM RC22176 (W0109-022), IBM Research Division, T. J. Watson Research Center.

A. Popescu-Belis. 2002. Meta-evaluation of evaluation metrics. Tentative synthesis on the evaluation exercise, May 30. Talk presented at Workshop on MT Evaluation: Human Evaluators Meet Automated Metrics (LREC'02).

A. Sågvall Hein, E. Forsbom, J. Tiedemann, P. Weijnitz, I. Almqvist, L.-J. Olsson, and S. Thaning. 2002. Scaling up an MT prototype for industrial use - databases and data flow. In *Proceedings from the 3rd International Conference on Language Resources and Evaluation (LREC'02),* pp. 1759-1766, Las Palmas de Gran Canaria, Spain.