Preprocessing and Normalization for Automatic Evaluation of Machine Translation

Gregor Leusch and **Nicola Ueffing** and **David Vilar** and **Hermann Ney**
Lehrstuhl für Informatik VI
RWTH Aachen University
D-52056 Aachen, Germany,
{leusch,ueffing,vilar,ney}@i6.informatik.rwth-aachen.de

Abstract

Evaluation measures for machine translation depend on several common methods, such as preprocessing, tokenization, handling of sentence boundaries, and the choice of a reference length. In this paper, we describe and review some new approaches to them and compare these to state-of-the-art methods. We experimentally look into their impact on four established evaluation measures. For this purpose, we study the correlation between automatic and human evaluation scores on three MT evaluation corpora. These experiments confirm that the tokenization method, the reference length selection scheme, and the use of sentence boundaries we introduce will increase the correlation between automatic and human evaluation scores. We find that ignoring case information and normalizing evaluator scores has a positive effect on the sentence level correlation as well.

1 Introduction

Machine translation (MT), as any other natural language processing (NLP) research subject, depends on the evaluation of its results. Unfortunately, human evaluation of MT system output is a time consuming and expensive task. This is why automatic evaluation is preferred to human evaluation in the research community.

Over the last years, a manifold of automatic evaluation measures has been proposed and studied. This underlines the importance, but also the complexity of finding a suitable evaluation measure for MT. We will give a short overview of some measures in section 2 of this paper.

Although most of these measures share similar ideas and foundation, we observe that researchers tend to approach problems common to several measures differently from each other. A noteworthy example here is the determination of a translation reference length.

In section 3, we will have a look onto structural similarities and differences among several measures, focussing on common steps. We will show that decisions taken about them can be as important to the outcome of an evaluation, as the choice of the evaluation measure itself.

To this end, we will study the performance of each error measure and setting by comparison with human evaluation on three different evaluation tasks in section 4. These experiments will show that sophisticated tokenization as well as adding sentence boundaries and a good choice for the reference lengths will improve correlation between automatic and human evaluation significantly. Case normalization and evaluator normalization are helpful only when evaluating on sentence level; system level evaluation is not affected by these methods.

After a discussion of these results in section 5, we will conclude this paper in section 6.

2 Automatic evaluation measures

The majority of MT evaluation approaches are based on the distance or similarity of MT candidate output to a set of reference translations, i.e. to sentences which are known to be correct. The lower this distance is, or the higher the similarity, the better the

Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 17–24, Ann Arbor, June 2005.

candidate translations are considered to be, and thus the better the MT system.

2.1 Evaluation measures studied

Out of the vast amount of measures, we will focus on the following measures that are widely used in research and in evaluation campaigns: WER, PER, BLEU, and NIST.

Let a test set consist of $k = 1, \ldots, K$ candidate sentences E_k generated by an MT system. For each candidate sentence E_k, we have a set of $r = 1, \ldots, R_k$ reference sentences $\widetilde{E}_{r,k}$. Let I_k denote the length, and I_k^* the reference length for each sentence E_k. We will explain in section 3.3 how the reference length is calculated.

With this, we write the total candidate length over the corpus as $\bar{I} := \sum_k I_k$, and the total reference length as $\bar{I}^* := \sum_k I_k^*$.

Let $n_{e_1^m,k}$ denote the count of the m-gram e_1^m within the candidate sentence E_k; similarly let $\tilde{n}_{e_1^m,r,k}$ denote the same count within the reference sentence $\widetilde{E}_{r,k}$. The total m-gram count over the corpus is then $\bar{n}_m := \sum_k \sum_{e_1^m \in E_k} n_{e_1^m,k}$.

2.1.1 WER

The word error rate is defined as the Levenshtein distance $d_L(E_k, \widetilde{E}_{r,k})$ between a candidate sentence E_k and a reference sentence $\widetilde{E}_{r,k}$, divided by the reference length I_k^* for normalization.

For a whole candidate corpus with multiple references, we define the WER to be:

$$\text{WER} := \frac{1}{\bar{I}^*} \sum_k \min_r d_L\big(E_k, \widetilde{E}_{r,k}\big)$$

Note that the WER of a single sentence can be calculated as the WER for a corpus of size $K = 1$.

2.1.2 PER

The position independent error rate (Tillmann et al., 1997) ignores the ordering of the words within a sentence. Independent of the word position, the minimum number of deletions, insertions, and substitutions to transform the candidate sentence into the reference sentence is calculated. Using the counts $n_{e,r}$, $\tilde{n}_{e,r,k}$ of a word e in the candidate sentence E_k, and the reference sentence $\widetilde{E}_{r,k}$, we can calculate this distance as

$$d_{\text{PER}}\big(E_k, \widetilde{E}_{r,k}\big) := \frac{1}{2}\Big(\big|I_k - \tilde{I}_k\big| + \sum_e \big|n_{e,k} - \tilde{n}_{e,r,k}\big|\Big)$$

This distance is then normalized into an error rate, the PER, as described in section 2.1.1.

A promising approach is to compare bigram or arbitrary m-gram count vectors instead of unigram count vectors only. This will take into account the ordering of the words within a sentence implicitly, although not as strong as the WER does.

2.1.3 BLEU

BLEU (Papineni et al., 2001) is a precision measure based on m-gram count vectors. The precision is modified such that multiple references are combined into a single m-gram count vector, $\tilde{n}_{e,k} := \max_r \tilde{n}_{e,r,k}$. Multiple occurrences of an m-gram in the candidate sentence are counted as correct only up to the maximum occurrence count within the reference sentences. Typically, $m = 1, \ldots, 4$.

To avoid a bias towards short candidate sentences consisting of "safe guesses" only, sentences shorter than the reference length will be penalized with a brevity penalty.

$$\text{BLEU} := lp_{\text{BLEU}} \cdot \underset{m}{gm} \left\{ \frac{1}{s_m + \bar{n}_m} \cdot \Big(s_m + \sum_k \sum_{e_1^m \in E_k} \min\big(n_{e_1^m,k}\,,\, \tilde{n}_{e_1^m,k}\big)\Big) \right\}$$

with the geometric mean gm and a brevity penalty

$$lp_{\text{BLEU}} := \min\left(1\,,\, \exp\Big(1 - \frac{\bar{I}^*}{\bar{I}}\Big)\right)$$

In the original BLEU definition, the smoothing term s_m is zero. To allow for sentence-wise evaluation, Lin and Och (2004) define the BLEU-S measure with $s_1 := 1$ and $s_{m>1} := 0$. We have adopted this technique for this study.

2.1.4 NIST

The NIST score (Doddington, 2002) extends the BLEU score by taking information weights of the m-grams into account. The NIST information weight is defined as

$$\mathit{Info}(e_1^m) := -\big(\log_2 \bar{\tilde{n}}_{e_1^m} - \log_2 \bar{\tilde{n}}_{e_1^{m-1}}\big)$$

with $\bar{\tilde{n}}_{e_1^m} := \sum_{k,r} \tilde{n}_{e_1^n,k,r}$.

Note that the weight of a phrase occurring in many references sentence for a candidate is considered to be lower than the weight of a phrase occurring only once!

The NIST score is the sum over all information counts of the co-occurring m-grams, summed up separately for each $m = 1, \ldots, 5$ and normalized by the total m-gram count.

$$\text{NIST} := lp_{\text{NIST}} \cdot \sum_m \left(\frac{1}{\bar{n}_m} \cdot \sum_k \sum_{e_1^m \in E_k} \min\left(n_{e_1^m,k}\,,\, \tilde{n}_{e_1^m,k}\right) \cdot \mathit{Info}(e_1^m) \right)$$

As in BLEU, there is a brevity penalty to avoid a bias towards short candidates:

$$lp_{\text{NIST}} := \exp\left(\beta \cdot \log_2^2 \min\left(1\,,\, \frac{\bar{I}}{\bar{I}^*}\right)\right)$$

$$\text{where } \beta := -\frac{\log_2 2}{\log_2^2 3}$$

Due to the information weights, the value of the NIST score depends highly on the selection of the reference corpus. This must be taken into account when comparing NIST scores of different evaluation campaigns.

2.2 Other measures

Lin and Och (2004) introduce a family of three measures named ROUGE. ROUGE-S is a skip-bigram F-measure. ROUGE-L and ROUGE-W are measures based on the length of the longest common subsequence of the sentences. ROUGE-S has a structure similar to the bigram PER presented here. We expect ROUGE-L and ROUGE-W to have similar properties to WER.

In (Leusch et al., 2003), we have described INVWER, a word error rate enhanced by block transposition edit operations. As structure and scores of INVWER are similar to WER, we have omitted INVWER experiments in this paper.

3 Preprocessing and normalization

Although the general idea is clear, there are still several details to be specified when implementing and using an automatic evaluation measure. We are going to investigate the following problems:

The first detail we have to state more precisely is the term "word" in the above formulae. A common approach for western languages is to consider spaces as separators of words. The role of punctuation marks in tokenization is arguable though. A punctuation mark can separate words, it can be part of a word, and it can be a word of its own. Equally it can be irrelevant at all for evaluation.

On the same lines it is to be specified whether we consider words to be equal if they differ only with respect to upper and lower case. For the IWSLT evaluation, (Paul et al., 2004) give an introduction to how the handling of punctuation and case information may affect automatic MT evaluation.

Also, a method to calculate the "reference length" must specified if there are multiple reference sentences of different length.

Since we want to compare automatic evaluation with human evaluation, we have to clarify some questions about assessing human evaluation as well: Large evaluation tasks are usually distributed to several human evaluators. To smooth evaluation noise, it is common practice to have each candidate sentence evaluated by at least two human judges independently. Therefore there are several evaluation scores for each candidate sentence. We require a single score for each system, though. Consequently, we have to specify how to combine the evaluator scores into sentence scores and then the sentence scores into a system score.

Different definitions of this will have a significant impact on automatic and human evaluation scores.

3.1 Tokenization and punctuation

The importance of punctuation as well as the strictness of punctuation rules depends on the language. In most western languages, correct punctuation can vastly improve the legibility of texts. Marks like full stop or comma separate words. Other marks like apostrophes and hyphens can be used to join words, forming new words by this. For example, the spelling "There's" is a contraction of "There is".

Similar phenomena can be found in other languages, although the set of critical characters may vary. Even when evaluating English translations, the candidate sentences may contain source language parts like proper names which should thus be treated according to the source language.

From the viewpoint of an automatic evaluation measure, we have to decide which units we would consider to be words of their own.

We have studied four tokenization methods. The simplest method is keeping the original sentences, and considering only spaces as word separators. Moreover, we can consider all punctuation marks to separate words but remove them completely then. The `mteval` tool (Papineni, 2002) improves this

Table 1: Tokenization methods studied

- Original candidate
  `Powell said: "We'd not be alone; that's for sure."`
- Remove punctuation
  `Powell said We d not be alone that s for sure`
- Tokenization of punctuation (`mteval`)
  `Powell said : " We'd not be alone ; that's for sure . "`
- Tokenization and treatment of abbreviations and contractions
  `Powell said : " we would not be alone ; that is for sure . "`

scheme by keeping all punctuation marks as separate words except for decimal points and hyphens joining composita. We have extended this scheme by implementing a treatment of common English contractions. Table 1 illustrates these methods.

3.2 Case sensitivity

In western languages, maintaining correct upper and lower case can improve the readability of a text. Unfortunately, though the case of a word depends on the word class, classification is not always unambiguous. What is more, the first word in a sentence is always written in upper case. This lowers the significance of case information in MT evaluation, as even a valid reordering of words between candidate and reference sentence may lead to conflicting cases. Consequently, we investigated if and how case information can be exploited for automatic evaluation.

3.3 Reference length

Each automatic evaluation measure we have taken into account depends on the calculation of a reference length: WER, PER, and ROUGE are normalized by it, whereas NIST or BLEU incorporate it for the determination of the brevity penalty. In MT evaluation practise, there are multiple reference sentences for each candidate sentence, with different lengths each. It is thus not intuitively clear what the "reference length" is.

A simple choice here is the average length of the reference sentences. Though this is modus operandi for NIST, it is problematic with brevity penalty or F-measure based scores, as even candidate sentences that are identical to a shorter-than-average reference sentence – which we would intuitively consider to be "optimal" – will then receive a sub-optimal score.

BLEU incorporates a different method for the determination of the reference length in its default implementation: Reference length here is the reference sentence length which is closest to the candidate length. If there is more than one the shortest of them is chosen.

For measures based on the comparison of single sentences such as WER, PER, and ROUGE, at least two more methods deserve consideration:

- The average length of the sentences with the lowest absolute distance or highest similarity to the candidate sentence. We call this method "average nearest-sentence length".
- The length of the sentence with the lowest relative error rate or the highest relative similarity. We call this method "best length". Note that when using this method, not the minimum absolute distance is used for the error rate, but the distance that leads to minimum relative error.

Other strategies studied by us, e.g. minimum length of the reference sentences, did not show any theoretical or experimental advantage over the methods mentioned here. Thus we will not discuss them in this paper.

3.4 Sentence boundaries

The position of a word within a sentence can be quite significant for the correctness of the sentence.

WER, INVWER, and ROUGE-L take into account the ordering explicitly. This is not the case with n-PER, BLEU, or NIST, although the positions of inner words are regarded implicitly by m-gram overlap. To model the position of words at the initial or the end of a sentence, one can enclose the sentence with artificial sentence boundary words. Although this is a common approach in language modelling, it has to our knowledge not yet been applied to MT evaluation.

3.5 Evaluator normalization

For human evaluation, it has to be specified how to handle evaluator bias, and how to combine sentence scores into system scores.

Regarding evaluator bias, even accurate evaluation guidelines will not prevent a measurable discrepancy between the scores assigned by different human evaluators.

The 2003 TIDES/MT evaluation may serve as an example here: Since the candidate sentences of

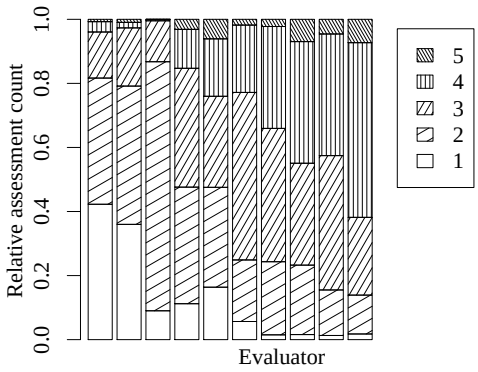

Figure 1: Distribution of adequacy assessments for each human evaluator. TIDES CE corpus.

the participating systems were randomly distributed among ten human evaluators, one would expect the assessed scores to be independent of the evaluator. Figure 1 indicates that this is indeed not the case, as the evaluators can clearly be distinguished by the amount of good and bad marks they assessed.

$(0, 1)$ evaluator normalization overcomes this bias: For each human evaluator the average sentence score given by him or her and its variance are calculated. These assignments are then normalized to $(0, 1)$ expectation and standard deviation (Doddington, 2003), separately for each evaluator.

Evaluator normalization should be unnecessary for system evaluation, as the evaluator biases tend to cancel out over the large amount of candidate sentences if the alignment of evaluators and systems is random enough. Moreover, with $(0, 1)$ normalization the calculated system scores are relative, not absolute scores. As such they can only be compared with scores out of the same evaluation.

Whereas the assessments by the human evaluators are given on the sentence level, our interest may lie on the evaluation of whole candidate systems. Depending on the number of assessments per candidate sentence, different combination methods for the sentence scores can be considered for this, e.g. mean or median. As our data consisted only of two or three human assessments per sentence, we have only applied the mean in our experiments.

It has to be defined how a system score is calculated from the sentence scores. All of the automatic evaluation measures implicitly weight the candidate sentences by their length. Consequently, we applied for the human evaluation scores a weighting by length on sentence level as well.

Table 2: Corpus statistics

	TIDES CE	TIDES AE	BTEC CE
Source language	Chinese	Arabic	Chinese
Target language	English	English	English
Sentences	919	663	500
Running words	25784	17763	3632
Punctuation marks	3760	2698	610
Ref. translations	4	4	16
Avg. ref. length	28.1	26.8	7.3
Candidate systems	7	6	11

4 Experimental results

To assess the impact of the mentioned preprocessing steps, we calculated scores for several automatic evaluation measures with varying preprocessing, reference length calculation, etc. on three evaluation test sets from international MT evaluation campaigns. We then compared these automatic evaluation results with human evaluation of adequacy and fluency by determining a correlation coefficient between human and automatic evaluation. We chose Pearson's r for this. Although all evaluation measures were calculated using length weighting, we did not do any weighting when calculating the sentence level correlation.

Regarding the m-gram PER, we had studied m-gram lengths of up to 8 both separately and in combination with shorter m-gram lengths in previous experiments. However, an m-gram length of greater than 4 did not show noteworthy correlation. For this, we will leave out these results in this paper.

For the sake of clarity, we will also leave out measures that behave very similarly to akin measures e.g. INVWER and WER, 2-PER and 1-PER, or BLEU and BLEU-S.

Since WER and PER are error measures, whereas BLEU and NIST are similarity measures, the correlation coefficients with human evaluation will have opposite signs. For convenience, we will look at the absolute coefficients only.

4.1 Corpora

From the 2003 TIDES evaluation campaign we included both the Chinese-English and the Arabic-English test corpus in our experiments. Both were provided with adequacy and fluency scores between 1 and 5 for seven and six candidate sets respectively.

As we wanted to perform experiments on a corpus with a larger amount of MT systems, we also included the IWSLT BTEC 2004 Chinese-English

evaluation (Akiba et al., 2004). We restricted our experiments to the eleven MT systems that had been trained on a common training corpus.

Corpus statistics can be found in table 2.

4.2 Experimental baseline

In our first experiment we studied the correlation of the different evaluation measures with human evaluation at "baseline" conditions. These included no sentence boundaries, but tokenization with treatment of abbreviations, see table 1. For sentence evaluation, conditions included evaluator normalization. Case information was removed. We used these settings in the other experiments, too, if not stated otherwise.

Figure 2 shows the correlation between automatic and human scores. On the TIDES corpora the system level correlation is particularly high, at a moderate sentence level correlation. We assume the latter is due to the poor sentence inter-annotator agreement on these corpora, which is then smoothed out on system level. On the BTEC corpus a high sentence level correlation accompanies a significantly lower system level correlation. Note that due to the much lower number of samples on the system level (e.g. 5 vs. 5500), small changes in the sentence level correlation are more likely to be significant than such changes on system level. We have verified these effects by inspecting the rank correlation on both levels, as well as by experiments on other corpora. Although these experiments support our findings, we have omitted results here for the sake of clarity.

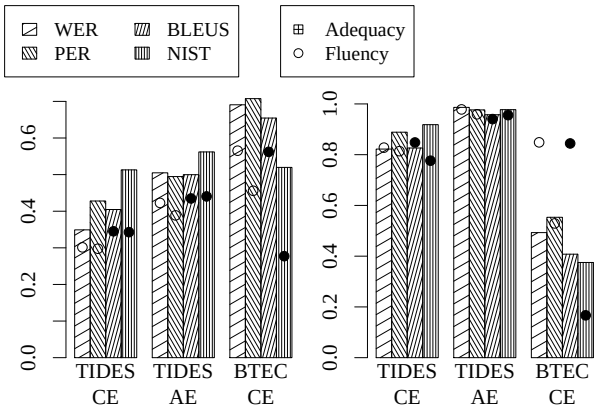

Figure 2: Pearson correlation coefficient between automatic and human evaluation. Bars indicate correlation with adequacy, circles with fluency score.
Left: sentence, **right:** system level correlation.

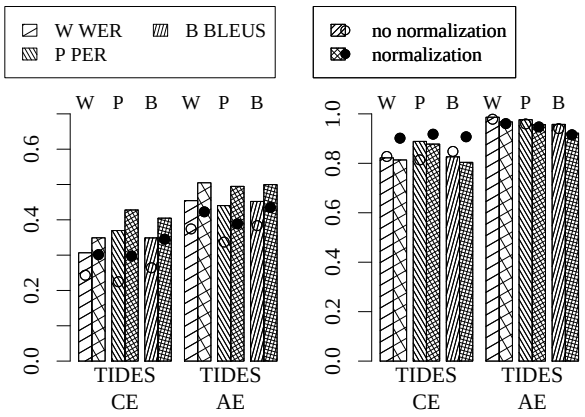

Figure 3: Effect of evaluator normalization.
Left: sentence, **right:** system level correlation.

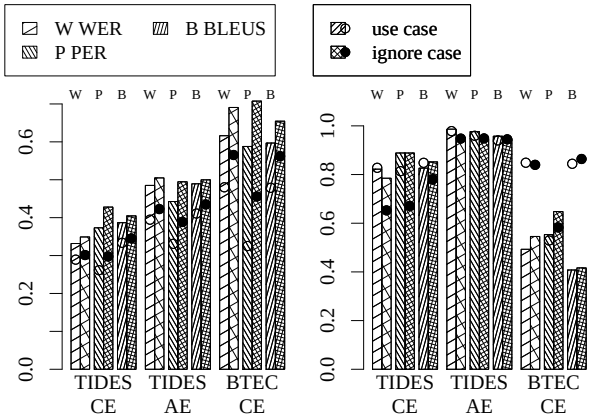

Figure 4: Effect of case normalization.
Left: sentence, **right:** system level correlation.

4.3 Evaluator normalization

We studied the effect of $(0, 1)$-normalization of scores assigned by human evaluators. The NIST measure showed a behavior very similar to that of the other measures and is thus left out in the graph. The correlation of all automatic measures both with fluency and with adequacy increases significantly at sentence level (figure 3). We do not notice a positive effect on system level, which confirms the assumption stated in section 3.5.

4.4 Tokenization and case normalization

The impact of case information was analyzed in our next experiment. Figure 4 (again without the NIST measure as it shows a similar behavior to the other measures) indicates that it is advisable to disregard case information when looking into adequacy on sentence level. Surprisingly, this also holds for

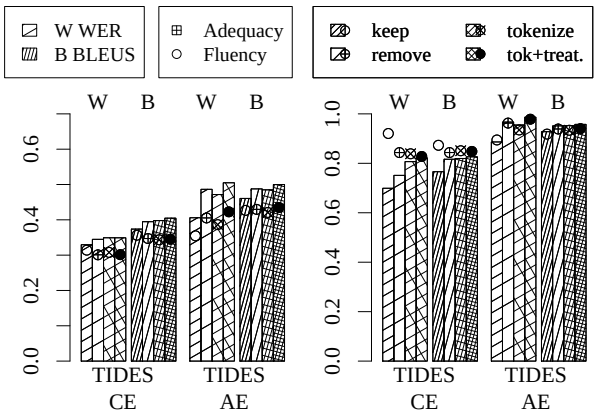

Figure 5: Effect of different tokenization steps.
Left: sentence, **right:** system level correlation.

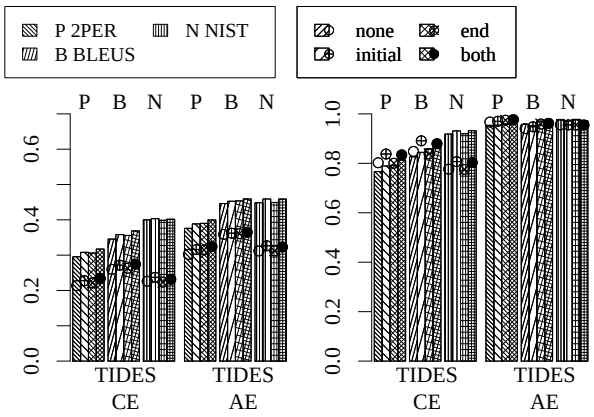

Figure 7: Effect of sentence boundaries.
Left: sentence, **right:** system level correlation.

fluency. We do no find a clear tendency on whether or not to regard case information at system level.

Figure 5 indicates that the way of handling punctuation we proposed does pay off when evaluating adequacy. For fluency our results were contradictory: A slight decrease on the Arabic-English corpus is accompanied by a slight decay on the Chinese-English corpus. We did not investigate the BTEC corpus here as most systems sticked to the tokenization guidelines for this evaluation.

4.5 Reference length

The dependency of evaluation measures on the selection of reference lengths is rarely covered in the literature. However, as we can see in figure 6, our experiments indicate a significant impact. The selected three methods here are the default for WER/PER, NIST, and BLEU, respectively. For the distance based evaluation measures, represented by WER here, taking the length of the sentence leading to the best score leads to the best correlation with both fluency and adequacy. Taking the average length instead seems to be the worst choice.

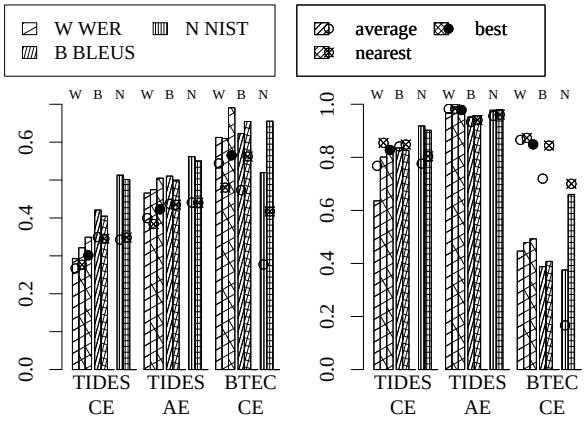

Figure 6: Effect of different reference lengths.
Left: sentence, **right:** system level correlation.

For brevity penalty based measures, the effect is not as clear: On both TIDES corpora there is no significant difference in correlation between using the average length and the nearest length. On the BTEC corpus, choosing the nearest sentence length leads to a significantly higher correlation than choosing the average length. We assume this is due to the high number of reference sentences on this corpus.

4.6 Sentence boundaries

As sentence boundaries will only influence m-gram count vector based measures, we have restricted our experiments to bigram PER, BLEU-S, and NIST here. Including sentence boundaries (figure 7) has a positive effect on correlation with fluency and adequacy for both bigram PER and BLEU-S. Sentence initials seem to be more important than sentence ends here. For the NIST measure, we do not find any significant effect.

5 Discussion

In a perfect MT world, any dependency of an evaluation on case information or tokenization should be inexistent, as MT systems already have to deal with both in the translation process, and could be designed to produce output according to evaluation campaign guidelines. Once all translation systems stick to the same specifications, no further preprocessing steps should be necessary.

In practice there will be some systems that step

out of line. If we then choose strict rules regarding case information and punctuation, automatic error measures will penalize these systems rather hard, whereas penalty is rather low if we choose lax ones.

In this situation case information will have a large effect on the correlation between automatic and human evaluation, depending on whether the involved candidate systems will have a good or a bad human evaluation. It is vital to keep this in mind when drawing conclusions here regarding system evaluation, despite the obvious importance of case information in natural languages.

These considerations also hold for the treatment of punctuation marks, as a special care should be unnecessary if all systems sticked to tokenization specifications. In practise, MT systems differ in the way they generate and handle punctuation marks. Therefore, appropriate preprocessing steps are advisable.

Our experiments suggest that sentence boundaries increase correlation between automatic scores and adequacy both on sentence and on system level. For fluency, the improvement is less significant, and mainly depends on the sentence initials.

For length penalty based measures, we have found that choosing the nearest sentence length yields the highest correlation with human evaluation. For distance based measures instead, it seems advisable to choose the sentence that leads to the best relative score as the one that determines the reference length.

6 Conclusion

We have described several MT evaluation measures. We have pointed out common preprocessing steps and auxiliary methods which have not been studied in detail so far in spite of their importance for the MT evaluation process. Particularly, we have introduced a novel method for determining the reference length of an evaluation candidate sentence, and a simple method to incorporate sentence boundary information to m-gram based evaluation measures.

We then have performed several experiments on these methods on three evaluation corpora. The results indicate that both our new reference length algorithm and the use of sentence boundaries improve the correlation of the studied automatic evaluation measures with human evaluation. Furthermore, we have learned that case information should be removed when performing automatic sentence evaluation. On sentence level, evaluator normalization can improve the correlation between automatic and human evaluation.

Acknowledgements

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG) under the project "Statistische Textübersetzung" (Ne572/5) and by the European Union under the integrated project TC-STAR – Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738).

References

Y. Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, and J. Tsujii. 2004. Overview of the IWSLT04 evaluation campaign. In *Proc. IWSLT*, pp. 1–12, Kyoto, Japan, September.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. ARPA Workshop on Human Language Technology*.

G. Doddington. 2003. NIST MT Evaluation Workshop. Personal communication, July.

G. Leusch, N. Ueffing, and H. Ney. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proc. MT Summit IX*, pp. 240–247, New Orleans, LA, September.

C. Y. Lin and F. J. Och. 2004. Orange: a method for evaluation automatic evaluation metrics for machine translation. In *Proc. COLING 2004*, pp. 501–507, Geneva, Switzerland, August.

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

K. A. Papineni. 2002. The NIST `mteval` scoring software. `http://www.itl.nist.gov/iad/894.01/tests/mt/resources/scoring.htm`.

M. Paul, H. Nakaiwa, and M. Federico. 2004. Towards innovative evaluation methodologies for speech translation. In *Working Notes of the NTCIR-4 Meeting*, volume 2, pp. 17–21.

C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated DP based search for statistical translation. In *European Conf. on Speech Communication and Technology*, pp. 2667–2670, Rhodes, Greece, September.