An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research

Sonja Nießen, Franz Josef Och, Gregor Leusch, Hermann Ney

Lehrstuhl für Informatik VI
RWTH Aachen – University of Technology
D-52056 Aachen, Germany
{niessen,och,ney}@informatik.rwth-aachen.de

Abstract

In this paper we present a tool for the evaluation of translation quality. First, the typical requirements of such a tool in the framework of machine translation (MT) research are discussed. We define evaluation criteria which are more adequate than pure edit distance and we describe how the measurement along these quality criteria is performed *semi-automatically* in a fast, convenient and above all consistent way using our tool and the corresponding graphical user interface.

1. Introduction

Research in machine translation suffers from the lack of suitable, consistent, and easy-to-use criteria for the evaluation of the experimental results. The question of how the performances of different translation systems on a certain corpus can be compared or how the effects of small changes in the system prototypes can be judged in a fast and cheap way is still open.

Efforts in the field of the evaluation of translation quality have focussed on measuring the suitability of a certain translation program as part of a distinct natural language processing task (White and Taylor, 1998; Sparck Jones and Galliers, 1996). Evaluation methods, which are 'ideal' in this respect would be too time-consuming to help the daily work of machine translation research.

2. Quality Criteria in MT Research

When researchers compare the performances of different translation systems or when they are interested in the effects of small changes in the system prototypes, they typically measure one or both of the following criteria:

- Word Error Rate (WER): The edit distance $d(t, r)$ (number of insertions, deletions and substitutions) between the produced translation t and one predefined reference translation r is calculated. The edit distance has the great advantage to be automatically computable, and as a consequence, the results are inexpensive to get and reproducible, because the underlying data and the algorithm are always the same.

  The great disadvantage of the WER is the fact that it depends fundamentally on the choice of the sample translation. In machine translation this criterion is used e.g. in (Vidal, 1997), and (Tillmann et al., 1997).

- Subjective Sentence Error Rate (SSER): The translations are scored by classification into a small number K of quality classes, ranging from "perfect" to "absolutely wrong". In comparison to the WER, this criterion is more liable and conveys more information, but to measure the SSER is expensive, as it is not computed automatically but is the result of labourous evaluation by human experts. Besides, the results depend highly on the persons performing the evaluation and hence, the comparability of results is not guaranteed. Another disadvantage is the fact that the length of the sentences is not taken into account: The score of the translation of a long sentence has the same impact on the overall result as the score of the translation of a one-word sentence. The SSER is used e.g. in (Nießen et al., 1998).

3. Semi-Automatic Evaluation

One of the characteristics of MT research is the fact that different prototypes of translation systems are tested *many times* on one distinct set of test sentences (for example for adjusting parameter settings or examining the effects of slight changes in system design). Sometimes the resulting translations differ only in a small number of words.

The idea now is to store an input sentence s together with all translations $\mathcal{T}(s) = t_1, \ldots, t_K$ that have already been manually evaluated together with their scores in a database $\mathcal{DB}$.

In addition, a suitable graphical user interface permits convenient manipulation of the database and provides means for calculating several kinds of statistics on it.

This approach and the resulting evaluation tool give us the following opportunities:

- *automatically* return the scores of translations that have already occured at least once. Hence, consistency of quality judgements over time is guaranteed (see 3.1.1.).

- facilitate the evaluation of new translations, that differ only slightly from previous ones (see 5.2.). This makes evaluation more efficient and helps maintenance of consistency.

- *extrapolate* scores for new translations by comparison with similar sentences in $\mathcal{DB}$ (see 3.1.1.).

- define new types of quality criteria (see 3.2. and 3.3.).

3.1. Definition of SSER

In our evaluation scheme, each translation t for an input sentence s is assigned a score $v(s, t)$ ranging from 0 points ("nonsense") to $K = 10$ points ("perfect"):

$0 \equiv$ nonsense.
$1 \equiv$ some aspects of contents are conveyed.
...
$5 \equiv$ understandable with major syntactic errors.
...
$9 \equiv$ ok. Only slight errors in register or style or minimal syntax errors.
K=10 $\equiv$ perfect.

After we have gathered experiences with the manual evaluations, the evaluators reported, that the chosen granularity was to high and that they would prefer a lower number of quality classes, say $K = 6$.

The SSER of a set of translations $t_1^n = t_1 \dots t_n$ for a test corpus $s_1^n = s_1 \dots s_n$ ranges from 0 to 100:

$$\mathrm{SSER}(s_1^n, t_1^n)[\%] = 100 \left(1 - \frac{1}{Kn} \sum_{i=1}^{n} v(s_i, t_i) \right) . \quad (1)$$

Note that this definition is based on the assumption that each individual score has the same weight, not taking the lengths of the scored sentence or the source language equivalent into account. Of course the sentence lengths are implicitely considered by the human evaluators.

3.1.1. Extrapolation of SSER

When a new set of translations for the test corpus s_1^n is generated, some of the pairs (s_i, t_i) have typically already been evaluated and their scores can be extracted from the database $\mathcal{DB}$. The remaining – really new – pairs are evaluated and added to $\mathcal{DB}$.

Additionally, we can extrapolate the score for a new translation t_i as follows: Provided that $\mathcal{DB}$ contains at least one translation for s_i, we compare t_i to all candidates $t_{i1}, \dots, t_{iK_i}$ in $\mathcal{T}(s_i)$ to calculate the minimum difference in terms of edit distance (0, if $(s_i, t_i) \in \mathcal{DB}$):

$$d(t_i, \mathcal{T}(s_i)) = \min_{t \in \mathcal{T}(s_i)} d(t_i, t) \quad (2)$$

and adopt the average score of the most similar candidates

$$\hat{t}(t_i, \mathcal{T}(s_i)) = \{ t \in \mathcal{T}(s_i) \mid d(t_i, t) = d(t_i, \mathcal{T}(s_i)) \} \quad (3)$$

to extrapolate the score of t_i:

$$\hat{v}(s_i, t_i, \mathcal{T}(s_i)) = \frac{1}{|\hat{t}(t_i, \mathcal{T}(s_i))|} \sum_{t \in \hat{t}(t_i, \mathcal{T}(s_i))} v(s_i, t) . \quad (4)$$

We define the extrapolated score as follows:

$$\tilde{v}(s, t) = \begin{cases} v(s, t) & \text{if } (s, t) \in \mathcal{DB} , \\ \hat{v}(s, t, \mathcal{T}(s)) & \text{otherwise} . \end{cases} \quad (5)$$

and define the extrapolated SSER eSSER by replacing $v(s_i, t_i)$ by $\tilde{v}(s_i, t_i)$ in definition (1).

As an indicator for the accuracy of this extrapolation, we compute the average normalized edit distance $\bar{d}(t_1^n)$:

$$\bar{d}(t_1^n) = \frac{1}{n} \sum_{i=1}^{n} \frac{d(t_i, \mathcal{T}(s_i))}{|s_i|} , \quad (6)$$

which depends on the rate of new translations as well as on the degree of similarity of these new hypotheses to the other candidates in the database. The above definition is a consequence of the definition of the SSER, which takes the quality judgements of all sentences as equally important.

3.2. Evaluation of information items

It remains unclear how to evaluate long sentences consisting of correct and wrong parts. To overcome this shortcoming of the SSER, we introduce the notion of "information items". Each input sentence s_i in the database is divided into segments representing the relevant information items to be conveyed. Then for each element of the set of information items for s_i, a candidate translation t_i is assigned either "ok" or one out of a predefined set of error classes. For our purposes we chose: "missing", "syntax", "meaning", and "other". The "information error rate" IER is the rate of information items not evaluated as "ok" for a set of translations t_1^n.

3.3. Definition of multi reference WER

We compute an "enhanced" WER as follows: a translation t_i is compared to *all* translations of s_i in $\mathcal{DB}$ that have been judged "perfect" (score 10) and the edit distance of t_i and the most similar sentence is used for the computation of the multi reference WER.

The idea of computing the difference to more than one reference has been used before (Alshawi et al., 1998). The advantage here is that the set of reference sentences comes for free as the database is enlarged. Besides, the new reference sentences produced by the translation systems under consideration are more adequate for the purpose of word-by-word comparison, because human translators tend to translate more or less freely, frequently resorting to synonyms and sentence restructuring.

4. The Database Format

We chose XML as format for the storage of evaluation databases. An example of a source sentence in German, segmented into two information items, with two corresponding translations together with their evaluation is shown below:

```
<database>
  ...
 <source>
  <s_sent>alles klar. danke schoen.</s_sent>
  <ielist>
   <iedef id="0">alles klar.</iedef>
   <iedef id="1">danke schoen.</iedef>
  </ielist>
  <targets>
   <tgt><t_sent>yes. thanks. fine.</t_sent>
    <eval val="6"/></tgt>
   <tgt><t_sent>okay thanks.</t_sent>
    <eval val="10"/>
        <ie id="0" val="ok"/>
        <ie id="1" val="ok"/></tgt>
   <tgt><t_sent>righto. thanks nice.</t_sent>
    <eval val="5"/></tgt>
  ...
  </targets>
 </source>
  ...
</database>
```

5. The Graphical User Interface (GUI)

We implemented a graphical user interface to facilitate the access to the database. For an overview, see Figure 1.

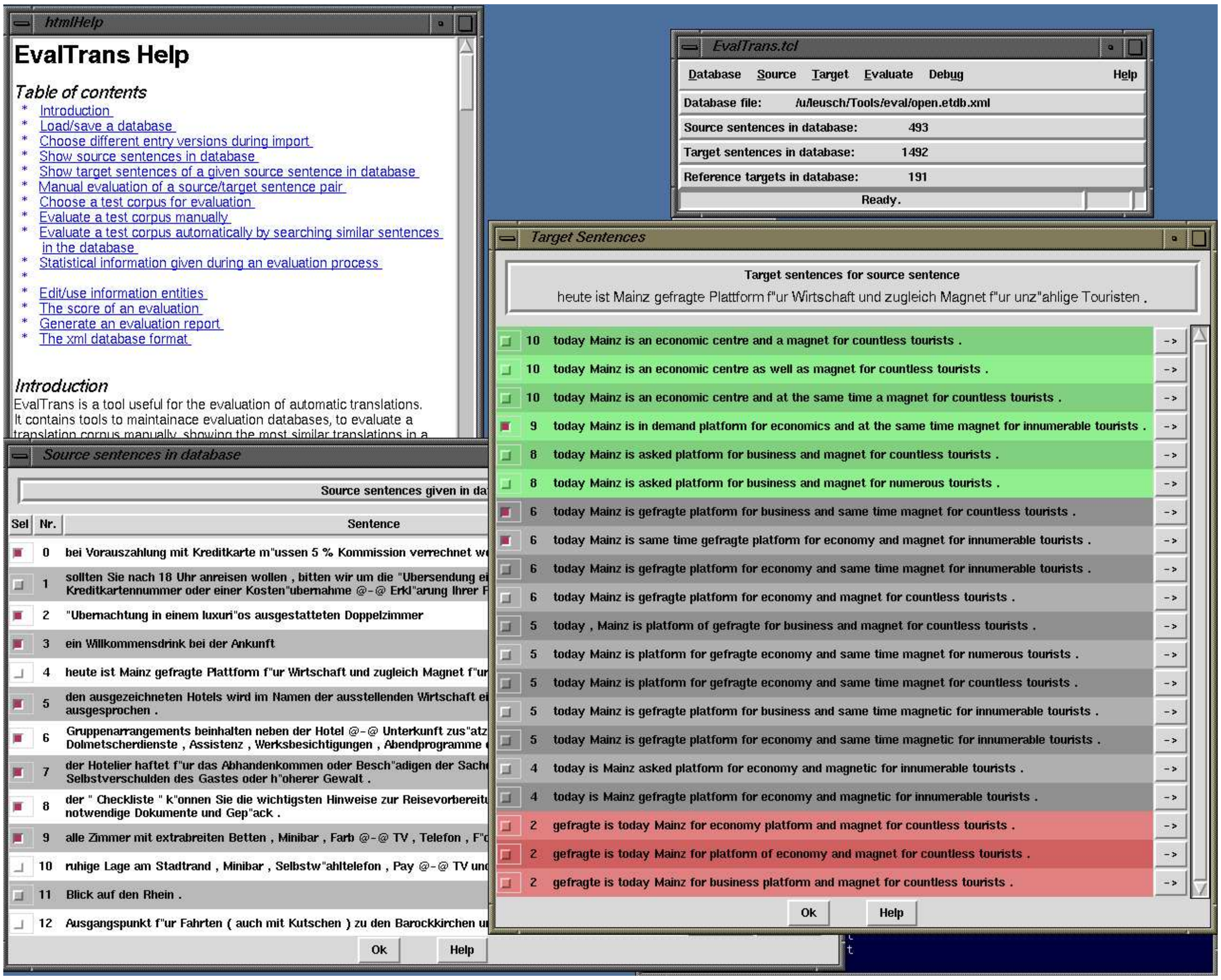

Figure 1: Overview of the GUI Layout.

The GUI offers the database manipulation operations import, export, selection, deletion and merging. The convenient segmentation of source sentences into information items is also supported. The implementation of a search method is planned. The interface also contains a help system based on hypertext.

The most important purpose of the GUI is on the one hand to display statistics about the status of the database and about a distinct set of candidate translations and on the other hand to facilitate the manual evaluation of new translations.

5.1. Displaying of Statistics

Three major kinds of statistics can be displayed:

1. For a selected source sentence s, compute the average number of correctly translated information items by sentences in $\mathcal{T}(s)$ (this conveys the “difficulty” of s). An example is shown in Figure 2.

2. For any subset of all scored and stored target sentences, display the average (absolute) extrapolation error (see section 6.2.1. and Figure 3).

3. For a given set of n pairs $(s_1, t_1) \dots (s_n, t_n)$, the following operations are possible: Print the eSSER, the average normalized edit distance $\bar{d}(t_1^n)$ and the average multi reference WER and for all pairs (s_i, t_i), print the extrapolated score $\tilde{v}(s_i, t_i)$, the minimal edit

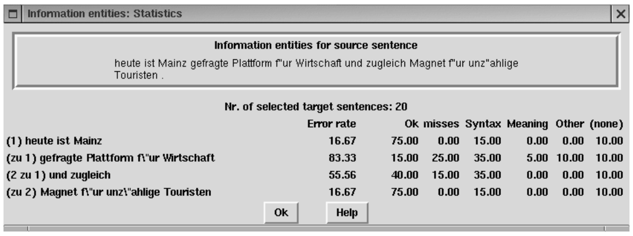

Figure 2: Statistics on information item error rate.

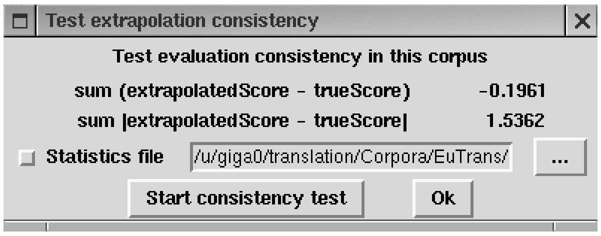

Figure 3: The average (absolute) extrapolation error.

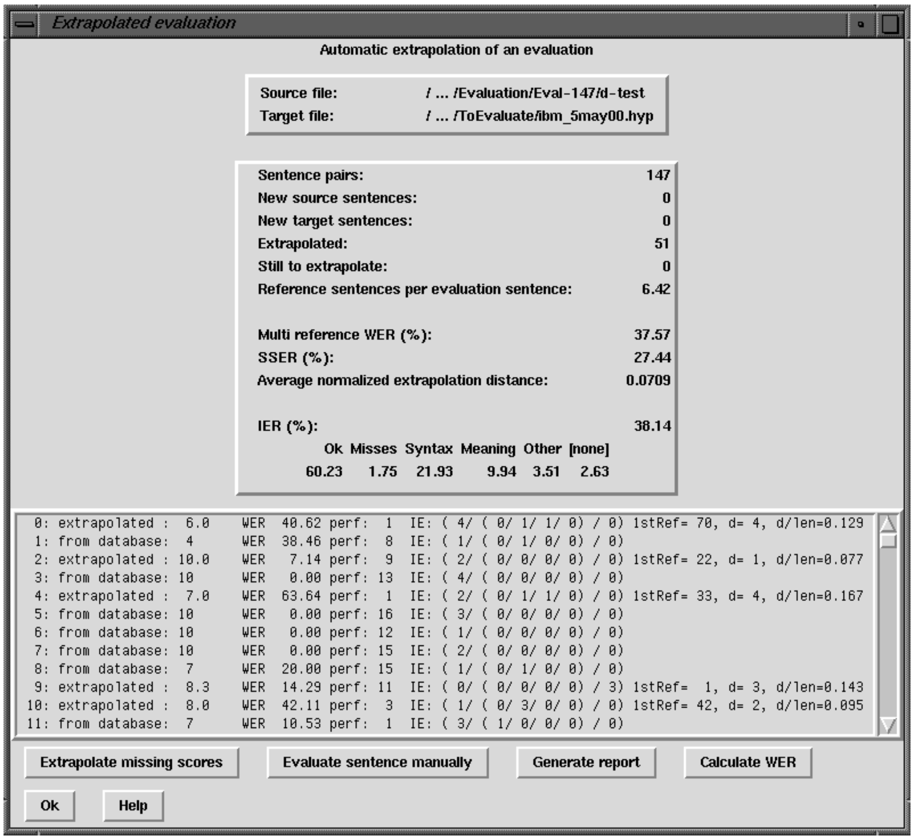

Figure 4: Statistics for a sample set of candidate translations.

distance $d(t_i, \mathcal{T}(s_i))$, the multi reference WER, and the number of information items translated correctly if (s_i, t_i) is already in the database. See Figure 4.

5.2. Manual Evaluation of new Translations

As can be seen in Figure 5, those candidate translations in $\mathcal{DB}$, that are most similar to t_i are highlighted. When moving the cursor over one of the candidates, all insertions, substitutions and deletions are marked in different colours. This facilitates the evaluation, as judgements can be made in comparison to other translations. The information items can be classified quickly by clicking on radio buttons for "ok" or one of the error classes.

6. Evaluation of the Tool

The machine translation research group at the department for Language Processing and Pattern Recognition at the University of Technology in Aachen constantly performs experiments to control the progress of the development of their translation systems. The Evaluation tool has yet been used for the evaluation of results on three different test sets, the first from the Verbmobil corpus (Wahlster, 1993) with spontaneously spoken dialogs in the domain of appointment scheduling and the other two from the Eu-Trans 2 Zeres corpus with texts in the touristic domain (see (Amengual et al., 1996) for a description of the first phase of this project). The corpus statistics and the range of the results on the test corpora for different translation methods in terms of SSER are briefly summarized in Table 1. The higher complexity of the EuTrans corpus (increased vocabulary size as well as smaller amount of training data and less constrained domain) results in higher SSER.

6.1. Efficiency of Manual Evaluation

The human evaluators who do the manual evaluation of the experimental results are students from the Department of English Language and Literature and the Department of Romance Languages.

They reported a substantial help for their work due to the graphical user interface. They also mentioned that the judgement of the information items not only caused an increased evaluation effort, but also helped getting a "feeling" for the quality of the translation under consideration. Highlighting of the most similar translation candidates and also marking the respective difference in terms of substitutions, insertions and deletions in different colours (see section 5.2. and Figure 5) helped speeding up the evaluation process substantially.

The evaluation of a new translation candidate needed approximately 30 to 60 seconds, depending on the length of the sentence, provided that the evaluators were already familiar with the *source* sentence.

6.2. Quality of Extrapolation

The accuracy of the extrapolation of the SSER depends on many factors, like complexity of the translation task, variability of the evaluated translations, degree to which the database is filled, i.e. number of translations per source sentence, etc. The average normalized edit distance $\bar{d}(t_1^n)$ is a measure for the reliability of the eSSER for a certain

Figure 5: Manual evaluation of a new translation candidate.

Table 1: Example of SSER and corpus statistics for various tasks.

		Verbmobil-147	EuTrans-closed	EuTrans-open
Words in Vocabulary	German	7 335	58 434	
	English	4 382	34 928	
Number of Sentences	Training	45 680	26 834	
	Test	147	100	100
range of the results in SSER		17% – 26%	57% – 76%	42%– 59%

set of new translations, whereas the methods described in subsection 6.2.1. allow for the computation of the expected extrapolation error on translations yet to be produced.

6.2.1. Leaving One Out validation (L1O)

As a measure for the reliability of the extrapolation of scores for new translation candidates, we compute the average absolute extrapolation error $|EE|(\mathcal{DB})$ ranging from 0 to 100. In the following definition, $|\mathcal{DB}|$ is the number of pairs (s,t) contained in $\mathcal{DB}$ (normalization constant):

$$|EE|(\mathcal{DB})[\%] = \frac{100}{K \cdot |\mathcal{DB}|} \sum_{\substack{(s,t) \\ \in \mathcal{DB}}} |v(s,t) - \hat{v}(s,t,\mathcal{T}(s) \setminus \{t\})|.$$

In words, the quantity conveys the following: For each target sentence t for a source sentence s, try to extrapolate the corresponding score from the *other* translation candidates (*leaving one out* scheme). The resulting estimate is compared to the real score of t. $|EE|(\mathcal{DB})$ gives the overall extrapolation error per sentence, i.e. a measurement for the reliability of the estimates for a distinct sentence. Note that the extrapolation process sometimes overestimates the quality of a translation, and sometimes the estimation is lower than the real score. It is for this reason that the eSSER on a set of n translation is more reliable than each extrapolated score of a distinct sentence t.

In Table 2, the results of the leaving one out validation on three different databases, representing different sets of test sentences, are summarized. The test sets are the same as summarized in Table 1. In Table 2, the column symbol "n" means "number of <u>different</u> source sentences" and "T/n" stands for "average number of target sentences per source sentence". Note that for the Verbmobil corpus n ist smaller than the number of sentences in the test corpus, because some sentences occur more than once.

Table 2: L1O Validation on different databases.

Database	n	T/n	$\|EE\|(\mathcal{DB})$
Verbmobil-147	144	41.3	11
Eutrans open	100	42.9	10
Eutrans closed	100	12.8	14

Figure 6 shows the development of the average absolute extrapolation error as the database is gradually filled. On the x-axis, the respective database version of the Verbmobil-147 evaluation database is shown (old versions can easily be retrieved, as the databases are under revision control).

6.2.2. Example hypotheses files

For 26 sets of translations (11 from the Verbmobil-147 test set, 6 from EuTrans open, 15 from EuTrans closed), we stored the eSSER and the corresponding $\bar{d}(t_1^n)$ just before evaluating them and compared the estimate to the real SSER afterwards (i.e. we computed the extrapolation error |SSER - eSSER|). The resulting diagram is shown in Figure 7. On the 26 files, the error |SSER - eSSER| was only 1.2% on average.

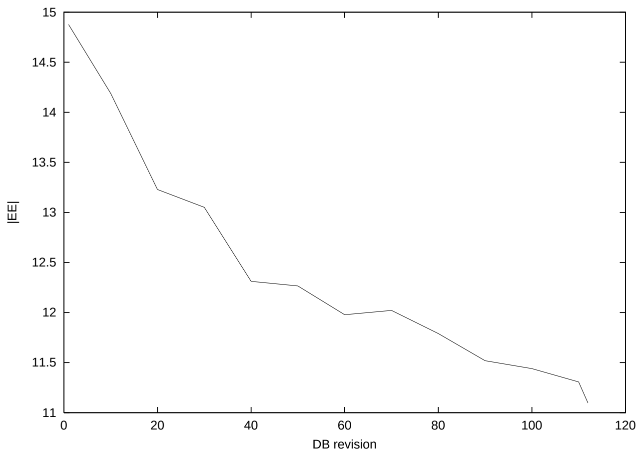

Figure 6: $|EE|(\mathcal{DB})$ versus revision number of $\mathcal{DB}$.

On average 29.5% of the sentences in the 26 sets described in this section had to be estimated, i.e. were not yet present in the database. This means that the tool saved at least 70% of the evaluation effort for the evaluation of these 26 translation hypotheses files!

6.3. Consistency of Results

The following experiment would convey information about the sensibility of the evaluation results against the so called "human factor", i.e. the question "how much would the SSER of a certain set of new candidates differ depending on which evaluator performs the evaluation and on his or her current mental constitution?": Randomly extract sentences with theirs scores from the database and make evaluators do the evaluation again. The resulting new score can be compared to the score formerly stored in the database. We have not performed this experiment so far.

6.4. Number of reference translations

In Table 3 the column symbol R/n means number of reference translations (score K) per source sentence. The EuTrans tests are more difficult than the test for Verbmobil. For this reason, and because less experiments have yet been run and thus less hypotheses have been evaluated for

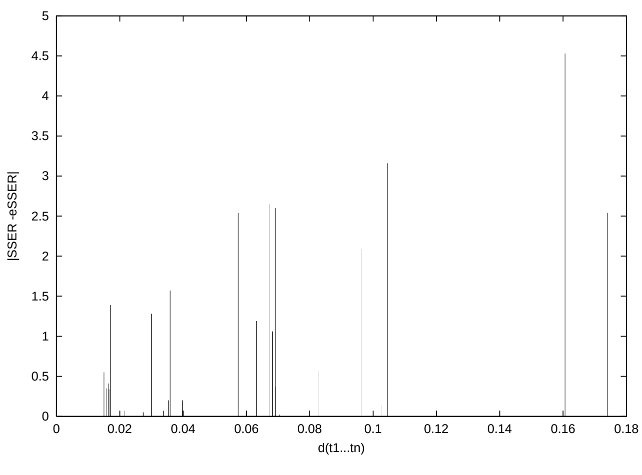

Figure 7: $\bar{d}(t_1^n)$ versus |SSER - eSSER|.

EuTrans-closed and especially for EuTrans-open, the number of reference translations is small compared to an average of 6 references for the Verbmobil test sentences.

Table 3: Number of reference translations.

Database	R/n
Verbmobil-147	6.0
EuTrans closed	1.3
EuTrans open	2.0

Figure 8 shows the development of the rate T/n of target sentences per source sentence and of the rate R/n of reference sentences per source sentence on the Verbmobil-147 database. Again, the x-axis represents increasing revision numbers.

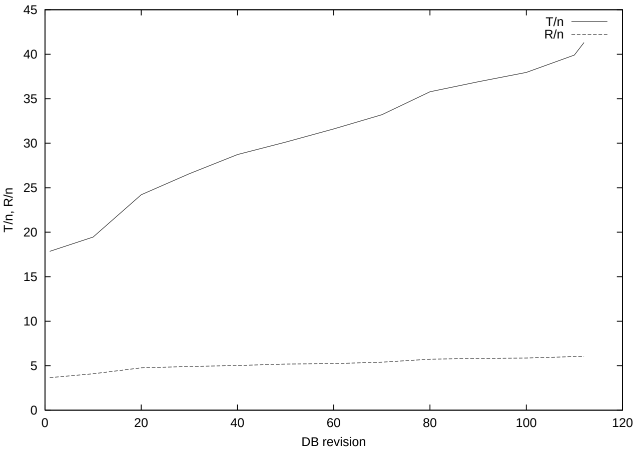

Figure 8: T/n and R/n versus revision number of $\mathcal{DB}$.

7. Further Applications and planned Improvements

We plan to facilitate the extraction of "difficult" source sentences in terms of average score and average rate of correctly translated information items of all candidate translations in $\mathcal{DB}$.

Some improvements of the current implementation are planned to make the tool more comfortable for the evaluators and to support consistency: More "natural" similarity measures than the traditional edit distance would allow for crossings in the two compared sentences. As a consequence, a more balanced selection of database entries to be offered as similar to the current translation candidate is possible. In future implementations, direct access to the information item evaluation of the most similar candidates will be provided to help maintaining the consistency between new and previous judgements.

A revised guideline for evaluators, containing qualitative descriptions of the classification criteria, is currently created.

The software will be made available for non-commercial purposes. If the reader is interested in using it, please feel free to send an email to one of the authors or to the MT research group (email adress: translation@i6.informatik.rwth-aachen.de).

Acknowledgement. This work was partly supported by the German Federal Ministry of Education, Science, Research and Technology under the Contract Number 01 IV 701 T4 (VERBMOBIL) and by the EUTRANS project by the European Community (ESPRIT project number 30268).

8. References

Alshawi, Hiyan, Srinivas Bangalore, and Shona Douglas, 1998. Automatic Acquisition of Hierarchical Transduction Models for Machine Translation. In *Proc. 36th Annual Conference of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*. Montréal, P.Q., Canada.

Amengual, J. C., J. M. Benedí, A. Castaño, A. Marzal, F. Prat, E. Vidal, J. M. Vilar, C. Delogu, A. di Carlo, H. Ney, and S. Vogel, 1996. Example-Based Understanding and Translation Systems (EuTrans): Final Report, Part I. Deliverable of ESPRIT project No. 20268.

Nießen, S., S. Vogel, H. Ney, and C. Tillmann, 1998. A DP based Search Algorithm for Statistical Machine Translation. In *Proc. 36th Annual Conference of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*. Montréal, P.Q., Canada.

Sparck Jones, Karen and Julia R. Galliers, 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Lecture notes in computer science. Berlin: Springer-Verlag.

Tillmann, C., S. Vogel, H. Ney, H. Sawaf, and A. Zubiaga, 1997. Accelerated DP based Search for Statistical Translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology, Rhodes, Greece*.

Vidal, E., 1997. Finite-State Speech-to-Speech Translation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Munich, Germany*.

Wahlster, W., 1993. Verbmobil: Translation of Face-to-Face Dialogs. In *Proceedings of the MT Summit IV*. Kobe, Japan.

White, John S. and Kathryn B. Taylor, 1998. A Task-Oriented Evaluation Metric for Machine Translation. In *Proc. First International Conference on Language Resources and Evaluation*. Granada, Spain.