# A Quantitative Method for Machine Translation Evaluation

**Jesús Tomás**
Escola Politècnica Superior de Gandia
Universitat Politècnica de València
jtomas@upv.es

**Josep Àngel Mas**
Departament d'Idiomes
Universitat Politècnica de València
jamas@idm.upv.es

**Francisco Casacuberta**
Institut Tecnològic d'Informàtica
Universitat Politècnica de València
fcn@iti.upv.es

## Abstract

Accurate evaluation of machine translation (MT) is an open problem. A brief survey of the current approach to tackle this problem is presented and a new proposal is introduced. This proposal attempts to measure the percentage of words, which should be modified at the output of an automatic translator in order to obtain a correct translation. To show the feasibility of the method we have assessed the most important Spanish-Catalan translators in comparing the results obtained by the various methods.

## 1 Introduction

Research in automatic translation lacks an appropriate, consistent and easy to use criterion for evaluating the results (White et al., 1994; Niessen et al., 2000). However, it turns out to be indispensable to have some tool that may allow us to compare two translation systems or to elicit how any variation of our system may affect the quality of the translations. This is important in the field of research as well as when a user has to choose between two or more translators.

The evaluation of a translation system shows a number of inherent difficulties. First of all we are dealing with a subjective process, which is even difficult to define.

This paper is circumscribed to the project SISHITRA (*SIStemas HÍbridos para la TRAducción valenciano-castellano* supported by the Spanish Government), whose aim is the construction of an automatic translator between Spanish and Catalan texts using hybrid methods (both deductive and inductive).

In the following section we discuss some of the most important translation quality metrics. After that, we introduce a semiautomatic methodology for MT evaluation and we show a tool to facilitate this kind of evaluation. Finally, we present the results obtained on the evaluation of several Spanish-Catalan translators.

## 2 Metrics in MT Evaluation

### 2.1 Automatic Evaluation Criteria

Within the scope of inductive translation, the use of objective metrics, which can be evaluated automatically, is quite frequent. These metrics take as their starting point a possible reference translation for each of the sentences we want to translate. This reference will be compared with the proposed sentences by the translation system. The most important metric systems are:

**Word Error Rate (WER):**
WER is the percentage of words, which are to be inserted, deleted or replaced in the translation in order to obtain the sentence of reference (Vidal, 1997; Tillmann et al., 1997). WER can be obtained automatically by using the editing distance between both sentences. This metric is computed efficiently and is reproducible (successive applications to the same data produce the same results). However, the main drawback is its dependency on the sentences of reference. There is an almost unlimited number of correct translations for one and the same sentence and, however, this metric considers only one to be correct.

**Sentence Error Rate (SER):**
SER indicates the percentage of sentences, whose translations have not matched in an exact manner those of reference. It shows similar advantages and shortcomings as WER.

Some variations on WER have been defined, which can also be obtained automatically:

**Multi reference WER (mWER):**
Identical approach to WER, but it considers several references for each sentence to be translated, i.e., for each sentence the editing distance will be calculated with regard to the various references and the smallest one is chosen (Niessen et al., 2000). It presents the drawback of requiring a great human effort before actually being able to use it. However, the effort is worthwhile, if it can be later used for hundreds of evaluations.

**BLEU Score:**
BLEU is an automatic metric designed by IBM, which uses several references (Papineni et al., 2002). The main problem of mWER is that all possible reference translations cannot be introduced. The BLEU score try to solve this problem by combining the available references. In a simplified manner we could say that it measures how many word sequences in the sentence under evaluation match the word sequences of some reference sentence. The BLEU score also includes a penalty for translations whose length differs significantly from that of the reference translation.

## 2.2 Subjective Evaluation Criteria

Other kinds of metrics have been developed, which require human intervention in order to obtain an evaluation. Among the most widely used we could stand out:

**Subjective Sentence Error Rate (SSER)**
Each sentence is scored from 0 to 10, according to its translation quality (Niessen et al., 2000).
An example of these categories is:

0 – nonsensical...
1 – some aspects of the content are conveyed
...
5 – comprehensible, but with important syntactic errors
...
9 – OK. Only slight style errors.
10 – perfect.

The biggest problem shown by this technique is its subjective nature. Two people who may evaluate the same experiment could obtain quite different results. To solve this problem several evaluations can be performed. Another drawback is that the different sentence lengths have not been taken into account. The score of a 100 word-long sentence has the same impact on the total score as that of a word-long sentence.

**Information Item Error Rate (IER)**
An unclear question is how to evaluate long sentences consisting of correct and wrong parts. IER attempts to find a solution to this question. In order to solve the problem the concept of "information items" is introduced. The sentences are divided into word segments. Each item of the sentence is marked with "OK", "error", "syntax", "meaning" or "others", as shown in the translation. The metric IER (Information Item Error Rate) can then be calculated as the percentage of badly translated items (not marked as "OK") (Niessen et al., 2000).

## 2.3 New Evaluation Criteria

Automatic metrics are especially useful, since their cost is practically null. However, they are very dependent on the used references. In some cases they can yield misleading results, for instance, if we want to compare an inductive translation system with some deductive one which, in principle, should produce translations of a similar quality. If we extract the references from the same source as the training material of the inductive translator, the inductive translator will have an advantage over the deductive translator, since it has learned to translate by using a vocabulary and structures that are similar to those appearing in the references.

| *acronym* | *name* | | *on* | *references* | *description* |
|---|---|---|---|---|---|
| WER | Word Error Rate | objective | word | 1 | % of words which are to be inserted, deleted or replaced in order to obtain the reference. |
| SER | Sentence Error Rate | | sent. | 1 | % of sentences different from reference. |
| mWER | Multi reference WER | | word | various | The same as WER, but with several reference sentences. |
| BLEU | Bilingual Evaluation Understudy | | sent. | various | The number of word groups that match the reference groups. |
| SSER | Subjective Sentence Error Rate | subjective | sent. | - | To each sentence a score from 0 to 10 is assigned. Later on, it is converted into %. |
| IER | Information Item Error Rate | | item | - | The sentence is segmented into information items. IER = % of badly translated items. |
| aWER | All references WER | | word | - | % of words to be inserted, deleted or replaced in order to obtain a correct translation. |
| aSER | All references SER | | sent. | - | % of incorrect sentences. |

**Table 1. Some metrics in MT evaluation**

The non-automatic evaluation metrics described above presents various constraints: When an SSER is used, it may be very difficult to decide the score to be assigned to one sentence. For example, if in one sentence a small syntactic error appears, we can assign an 8. If in the following sentence two similar errors appear, what score should we assign? The same or half the score? To solve these kinds of matters, IER introduces the concept of "information item". This proposal has the drawback of being quite costly, both during the initial stage of deciding the word segments which form each item as well as when classifying the correction for each item. After having seen the previous drawbacks the following metric has been introduced:

**All references WER (aWER):**
It measures the number of words, which are to be inserted, deleted or replaced in the sentence under evaluation in order to obtain a correct translation. It can also be seen as a particular case of the mWER, but taking for granted that all the possible references are at our disposal. Since it is impossible to have *a priori* all possible references, the evaluator will be able to propose new references, if needed. The evaluation process can be carried out very quickly, if one takes as the starting point the result obtained by the WER or the mWER. The idea consists of visualising the incorrect words detected by one of these methods (editing operations). The evaluator just needs to indicate whether each of the marked items is an actual error or whether it can rather be considered as an alternative translation

This metric resembles very much the one proposed in (Brown et al, 1990). That work suggested for measuring the translation quality counting the number of times an evaluator would have to press the keyboard keys in order to make the proposed sentence correct.

**All references Sentence Error Rate (aSER):**
The SER metric presents the drawback of working with only one reference. Therefore, it does not really measure the number of wrong sentences, but rather those that do not match exactly the reference. For this reason we thought it would be interesting to introduce a metric that could indicate the percentage of sentences whose

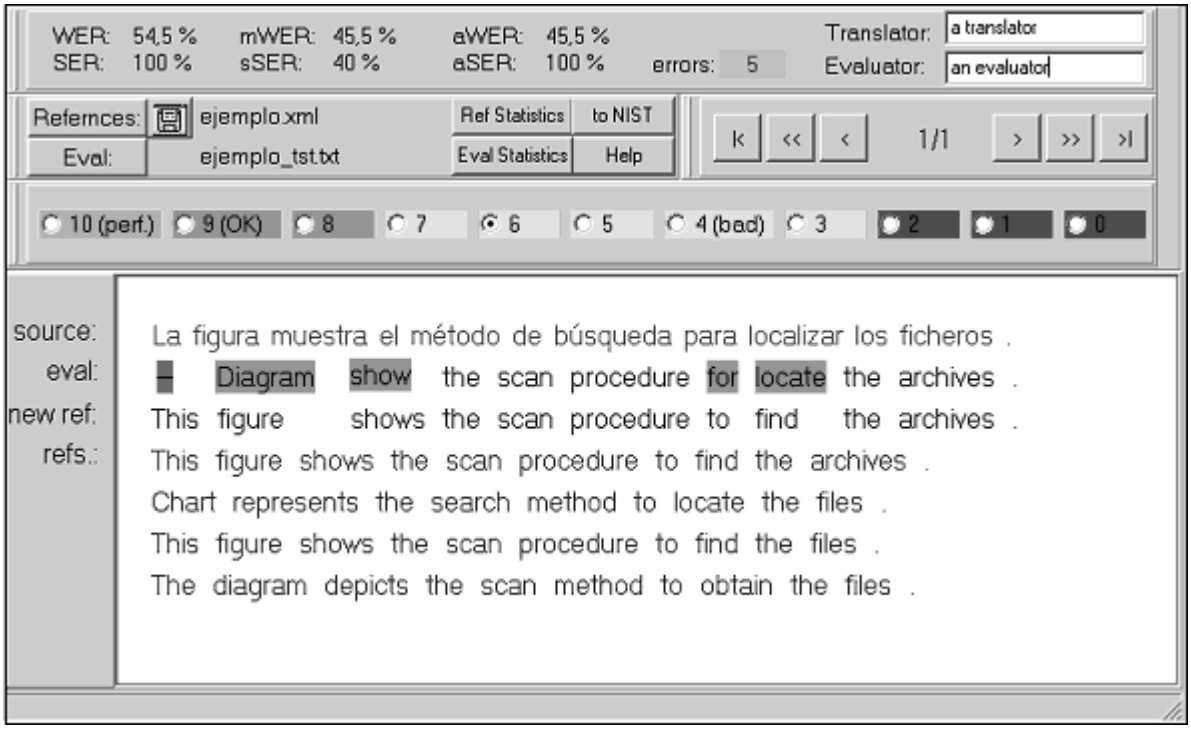


**Figure 1. The Graphic User Interface. The system highlights the non-matching words between the evaluation sentence and the nearest reference.**

translations are incorrect. This metric can be obtained as a by-product of the aWER.

## 3 Evaluation Tool for MT

In order to facilitate the evaluation of automatic translators a graphic user interface has been implemented. The metrics provided by the program are: WER, mWER, aWER, SER, SSER and aSER. Figure 1 shows how it is displayed.

Next, the way the program works is described:

On the editing window from top to bottom the following items are displayed: the source sentence, the sentence to be evaluated, the new sentences proposed by the user, the four most similar references to the sentence under evaluation (according to editing distance). The new sentence proposed by the user will be in principle the same as that of the most similar reference. In the sentence being evaluated using different colours, depending on whether they are considered insertions, replacements or deletions, the words that may be wrong are highlighted.

The user can click with the mouse on those words that may be considered correct. As a result, this action will modify the new reference. In the example (figure 1), if the user clicks on the highlighted words *"-"*, *"Diagram"* and *"locate"*, he will obtain the new reference *"Diagram shows the scan procedure to locate the archives."*. This new reference reduces the editing distance from 5 to 2. The user will also be able to click directly on some word of new reference to modify it. The aim of this is to allow the evaluator the introduction of any new reference which may be a correct translation of the source sentence and which, furthermore, may resemble most closely the sentence being evaluated.

This tool can be obtained for free on (http://ttt.gan.upv.es/~jtomas/eval), both in the Linux version as wells as in Windows.

### 3.1 Evaluation Database Format

A format in XML has been defined to store the reference files. For each evaluation sentence we store: the source sentence, the target reference sentences and the target sentences proposed by the different MT with their subjective

evaluations. Should during an aWER evaluation a new reference be proposed, this one is also stored. An example of a file with a sentence under evaluation is shown as follows:

```
<evalTrans>
<sentence>
  <source>
    La figura muestra el método.
  </source>
  <eval translator="first reference">
    <target>
      This figure shows the procedure.
    </target>
  </eval>
  <eval translator="multi reference">
    <target>
      This figure shows the method.
    </target>
  </eval>
  <eval translator="Statistical"
   evaluator="JM" sser="8" awer="1/5">
    <target>
      Chart represent the method.
    </target>
    <newRef>
      Chart represents the method.
    </newRef>
  </eval>
</sentence>
...
</evalTrans>
```

## 4 Example of Evaluation

### 4.1 Spanish-Catalan Translators

The tool described in the previous section has been applied to the most important Spanish-Catalan translators.

The Catalan language receives more or less intense institutional support in all territories of the Spanish state, where it is co-official with Spanish (Balearic Islands, Catalonia and Valencian Community). This makes it compulsory from an administrative standpoint to publish a bilingual edition of all official documents. For that purpose the use of a Machine Translator becomes almost indispensable.

But the official scope is not the only one where we can find the need to write bilingual documents in a short period of time. The most obvious example can be the bilingual edition of some newspapers, such as *El País* or *El Periódico de Catalunya*, both in their editions for the autonomous community of Catalonia.

In the following section there is a brief description of each of the programs we have reviewed:

***Salt***: an automatic translation program of the Valencian local government, which also includes a text corrector. It can be downloaded for free from http://www.cultgva.es. It has an interactive option for solving doubts (subjective ambiguity resolution) and is executed with the OS Microsoft Windows.

***Incyta***: the translation business web-site Incyta (http://www.incyta.com) was adding at the time of this evaluation example review a free on-line automatic translator for short texts.

***Internostrum***: an on-line automatic translation program, available at http://www.torsimany.ua.es, designed by the Language and Computational Systems Department of the University of Alicante. It marks the doubtful words or segments as a review helping aid. It uses finite-state technology (Canals et al., 2001).

***Statistical***: An experimental translator developed at the Computer Technology Institute of the Polytechnic University of Valencia. All components have been inferred automatically from training pairs using statistical methods (Tomás & Casacuberta, 2001). It is accessible at http://ttt.gan.upv.es/~jtomas/trad.

### 4.2 Setting up the evaluation experiment

In order to carry out our evaluation, we have translated 120 sentences (2456 words) with the different MT. These sentences have been taken from different media: a newspaper, a technical manual, legal text... The references used by the WER were also taken from the Catalan version of the same documents. In mWER and in BLEU we used three additional references. These new references have been introduced by a human translator modifying the initial reference.

Before applying the metrics shown in point 2, a human expert carries out a detailed analysis in order to establish the quality of the translations. The experiment consists of sorting out the four outputs obtained by each translator for each test sentence, according to its quality. If the expert does no find any quality difference between the

| *Translator* | first | second | thrid | fourth |
|---|---|---|---|---|
| Salt | 69% | 13% | 13% | 4% |
| Incyta | 63% | 11% | 13% | 13% |
| Statistical | 60% | 13% | 7% | 20% |
| Internostrum | 48% | 12% | 20% | 20% |

**Table 2. Comparative classification sentence by sentence.**

sentences proposed by two translators, he assigns the same rank to them. Table 2 shows the results obtained. After this sentence by sentence analysis, the expert concludes that *Salt* is the better translator, followed closely by *Incyta*. *Statistical* is in an intermediate position and the worst is *Internostrum*.

### 4.3 Results

The results of our experiment can be observed in Figure 2. Table 3 shows the evaluation time for the 120 sentences. The first thing we can point out is that the *Salt* translator obtains the best results from all used metrics and *Internostrum* is the worst of all metrics. The other two translators obtain different results depending on the used method. Next we will discuss the results obtained by the different methods:

The **WER** metric shows a strong dependence on the used reference. If the translator employs a similar style or vocabulary with regard to those of the reference, it clearly achieves better results. This fact determines that the obtained results do not show faithfully the quality of the translations. Specifically, for *Incyta* it obtains bad results, although that does not coincide with the conclusions of the expert.

The main advantage of this method is that it is a totally automatic measurement without any evaluation cost. These conclusions can also be extended to the SER.

**mWER** solves in part the problem posed by the WER. To attempt to introduce *a priori* all possible translations turns out to be impossible, so that it has to choose a subset of these giving thus the method a certain subjective nature. In the case of our evaluation, the references were introduced by using certain dialectal variants. That worked slightly against some automatic translator, which preferred some other dialectal variants.

The **BLEU** metric tries to combine the available references in order to improve the mWER metric. In our experiment the use of several references, in mWER and BLEU, does not solve the deficiency of WER. It continues being most detrimental to *Incyta*.

The use of the mWER and BLEU required a great initial effort, when the references were written, by even choosing only three new references for each translation. However, these methods had a big advantage: each evaluation is done without any additional cost.

When we applied the **SSER**, we faced the following dilemma: Which criteria should we use for applying the scoring scale? We decided that the latter had to be related with the global understanding of the sentence and the number of errors in correspondence with the sentence length. Since this criterion is not made explicit in the method the choice of a different criterion would have produced very diverse results.

Regarding the evaluation effort, it was the most costly method. In order to evaluate each sentence it was necessary to read and understand both the source sentence and the target sentence to try to score at the end the translation.

The **aWER** metric breaks with the dependence on the used references, which displayed the WER, mWER and BLEU. Moreover, it turned out to be much more objective and clearer to apply than the SSER. The metric achieved by this method provides us with clear and intuitive information. If we use the *Salt* translator we will have to correct 3% of the words in order to obtain a correct translation. Interpret the metrics supplied by the other methods it becomes unavoidable to know the conditions under which the evaluation has been carried out (references used, criteria ...).

The evaluation effort for the aWER is significantly less than the mWER and the SSER.

| *Translator* | WER | | mWER | | aWER | | SER | | SSER | | BLEU | | aSER | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Salt | 9.9 | *1* | 6.6 | *1* | 3.0 | *1* | 68.3 | *1* | 10.3 | *1* | 0.866 | *1* | 40.0 | *1* |
| Incyta | 10.9 | *2* | 7.6 | *2* | 3.1 | *1* | 74.2 | *2* | 11.2 | *1* | 0.855 | *2* | 41.7 | *1* |
| Statistical | 10.7 | *2* | 7.8 | *2* | 3.8 | *2* | 70.8 | *2* | 12.8 | *2* | 0.857 | *2* | 45.8 | *2* |
| Internostrum | 11.9 | *3* | 8.5 | *3* | 4.9 | *3* | 80.0 | *3* | 15.8 | *3* | 0.837 | *3* | 58.3 | *3* |

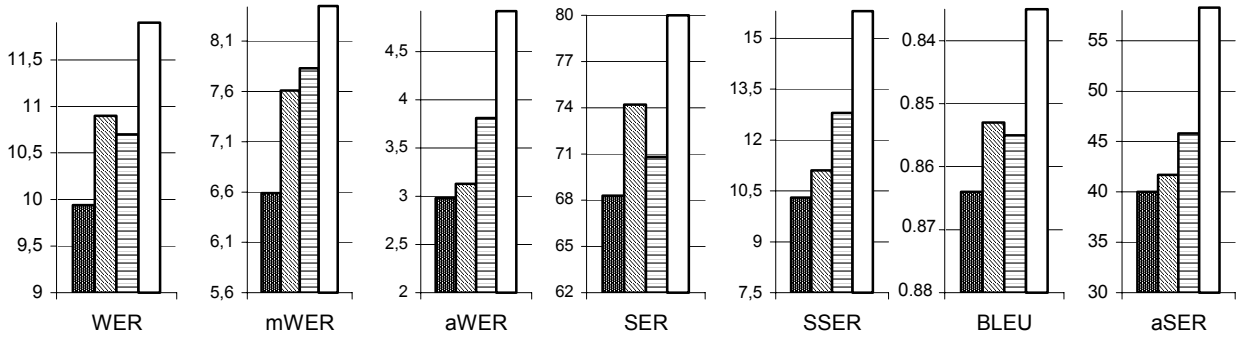


**Figure 2. Comparative evaluation results using 7 different metrics for the 4 Spanish-Catalan translators. In order to interpret quickly the results obtained in each metric, we have classified each translator using the following ranking: 1- better 2- intermediate 3- worse.**

| | mWER / BLEU | SSER | aWER / aSER |
|---|---|---|---|
| Set-up time* | 210 | 0 | 0 |
| Internostrum | 0 | 70 | 40 |
| Salt | 0 | 60 | 25 |
| Incyta | 0 | 55 | 30 |
| Statistical | 0 | 60 | 25 |
| **Total:** | 210 | 245 | 120 |

**Table 3. Comparative evaluation time (minutes) of the 120 sentences using the different metrics. *Time spent to introduce the proposed references.**

The discussion on the aWER method can be extended to the aSER.

Considering the expert evaluation, the subjective metrics reflect better the quality of the evaluated translations than the automatic ones. The *Incyta* translator works quite appropriately, but it proposes translations that deviate from the references. Thus, the automatic measures (WER, mWER and BLEU), based on these references, do not evaluate correctly this translator. On the other hand, the *Statistical* Translator works worse, even though its translations are more similar to the references. It is an example-based translator, and the training and test sentences have been obtained from the same sources. This can benefit the evaluation of the *Statistical* translator using automatic measures.

## 5 Conclusions

In this paper we present a criterion (aWER) for the evaluation of translation systems. The evaluation of the translations can be carried out quickly thanks to the use of a computer tool developed for this purpose.

We have compared this criterion with other criteria (WER, mWER, SER, BLEU and SSER) using the translations obtained by several Spanish-Catalan translators. It is our understanding that automatic measures (WER, mWER and BLEU) do not evaluate correctly the translators (specifically, they affect *Incyta* negatively).

The scores produced by human experts (SSER and aWER) are the metrics that best capture the translation quality among the different systems. As its most important aWER feature we would stand out that, in spite of being a subjective method which requires the intervention of a human evaluator, the latter will not have to take too subjective decisions.

We believe that the aWER tool could be used in another domain, for the evaluation of other natural language processing systems, e.g. summarizing systems.

In a future our aim is to add to this comparative study other score methods, in addition to comparing the variability introduced by different human evaluators in each of the methods.

## Acknowledgement

This work was partially funded by the Spanish CICYT under grant TIC2000-1599-C02 and the IST Programme of the European Union under grant IST-2001-32091. The authors wish to thank the anonymous reviewers for their criticism and suggestions.

## References

Brown, P. F., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, & P. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics* 16(2).

Canals-Marote, R., A. Esteve-Guillén, A. Garrido-Alenda, M.I. Guardiola-Savall, A. Iturraspe-Bellver, S. Montserrat-Buendia, S. Ortiz-Rojas, H. Pastor-Pina, P.M. Pérez-Antón, M.L. Forcada. 2001. The Spanish-Catalan machine translation system interNOSTRUM. In *Proceedings of the Machine Translation Summit VIII*. Santiago de Compostela, Spain.

Niessen, S., 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*, Athens, Greece.

Papineni, K.A., S. Roukos, T. Ward, W.J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the* 40th *Annual Meeting of the Association for Computational Linguistics* (ACL), Philadelphia.

Tomás, J., F. Casacuberta. 2001. Monotone Statistical Translation using Word Groups. In *Proceedings of the Machine Translation Summit VIII*. Santiago de Compostela, Spain.

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.

White, J., T. O'Connell, F. O'Mara. 1994. The DARPA Machine Translation Evaluation Methodologies: Evolution, Lessons and Future Approaches. In *Proceedings of the first Conference of the Association for Machine Translation in the Americas.* Columbia, USA.