Towards a Systematic Evaluation of Machine Translation Systems

Melchior Amgarten and Dominic A. Merz
University of Zürich
Department of Computer Science, Computational Linguistics
Winterthurerstr. 190, CH-8057 Zürich
mamgar@cl.unizh.ch · merz@ifi.unizh.ch

Abstract

There are still a lot of uncertainties about the evaluation of machine translation (MT) systems. In this report, we try to establish a catalog of criteria which we consider to be crucial for a systematic assessment of MT systems; the paper is divided into three parts: technical, linguistic and entrepreneurial criteria. Furthermore, we give hints on how to organize an evaluation project.

1 Introduction

It is not easy to decide on the right criteria for evaluating machine translation (MT) systems (cf. [Kin96]), but it is even more difficult to find precise measures which objectively describe what a good system consists of, or to find a suitable evaluation setup. Before looking at evaluation criteria let us consider different motivations for evaluation.

Basically, two groups might be interested in criteria to evaluate MT systems, each of which has its own preferences.

Firstly, a *user* might ask the following questions:

- Which of the MT systems available can provide me with the best translation quality?
- From which system do I gain the most profit for my company?

Secondly, a *system developer* might be interested in answers to:

- Which features of my MT system are already correctly implemented?
- Where in my system do I have to make corrections?
- In which direction do I have to continue developing my system?

We have collected a catalog of criteria for MT evaluation which we present in the following sections. In this catalog, we have tried to take into consideration the interests of both users and system developers.

2 The Different Criteria

We have divided the catalog into three parts: technical, linguistic and entrepreneurial criteria.

2.1 Technical Criteria

The technical criteria are important for the system's performance as such, but they are not directly connected with the translation output. The following list introduces six technical criteria which we have found to be central:

- speed
- user friendliness
- interface
- extensibility
- robustness
- maintenance

It goes without saying that a good MT system has to be fast. However, to evaluate the **speed** of a system you should measure the time of the translation process as a whole rather than the sheer translation time. This would also include the time spent for pre-editing, as well as the time spent for post-editing. In order to get a good result, the translation system should be configured well before the beginning of the translation, i.e. the subject areas, if available, should be chosen according to the topic of the text. This is the only way to provide equal opportunities, because otherwise the various results are influenced too much by the differences in configuration.

To evaluate the **user friendliness** and the **interface** you will depend on the help of volunteers. You should not evaluate these criteria yourself because the learning process you go through when testing one system after another will affect your evaluation results. Still, only persons who have used the system can tell you how easy or how hard it is to work with. You may question them, for instance, about the handling during a translation session, whether or not the system is clearly organized, if it is easy to acquire the ability to work with it, or whether the manuals were helpful in the case of some support needed.

Practical work with MT systems clearly shows that **extensibility** is an important point. Some systems allow you to add additional features like special dictionaries for specific subject areas, and it would be desirable to have the possibility to add a grammar checker for the target text. The importance of the extensibility of the lexicon and the grammar will be pointed out in the following subsection.

Also, you should test the **robustness** of a system by trying to cause it to crash. The system has to be solid and robust enough to cope with problems as, for example, sentences too long for translation, completely ungrammatical sentences in the source text, words not found in the lexicon or system out of memory.

Take into account that your system needs **maintenance**, on the one hand provided by yourself, on the other hand offered by the selling company. The system scores extra points, if, for instance, the provider will support you, if there is a hotline established, or if you are automatically provided with updates.

2.2 Linguistic Criteria

The crucial points for an evaluation of MT systems are the linguistic criteria, especially the quality of their translation output, but also their lexicon and the grammar implemented in them. Unfortunately, an evaluation according to these criteria is the most difficult.

For the **translation quality**, we have gathered a list of criteria which we introduce only informally — in section 3 we will have a closer look at ways to assess them:

Accuracy: An accurate translation stays close to the source text; it tries to follow the text and sentence structure of the source text as far as possible. Accuracy is often called fidelity to indicate the closeness to the original. Accuracy needs to be traded against adequacy.

Adequacy: An adequate translation aims at a useful and clear target text. In many cases you cannot translate a text word by word. For instance: the beginning phrase of a typical German fairy-tale, is *Es war einmal*. A word by word translation will produce *It was once*, whereas an English fairy-tale idiomatically starts with *Once upon a time*. Only the latter is an adequate translation for the German phrase.

Intelligibility: A target text is intelligible if the reader is able to find in it all and only the information of the source text.

Readability: The translation should be as easy to read as a text originally written in the target language. That means that sentences have to be (close to) grammatical and that the text be coherent.

Style: This criterion is especially hard for MT systems. Nevertheless, it is the goal of a good translation to adapt each translation to an appropriate style in the target language.

The translation quality is of course influenced by the organisation of the lexicon and the architecture of the grammar — let us now examine these essential resources.

The term **lexicon** refers to the system's translation dictionary. Not merely the size of the lexicon is important, but also a useful organisation, e.g. semantic features or a hierarchy of different subject areas. In addition to single words, the lexicon should contain complete phrases. One of the most important features is the extensibility of its lexicon. No matter how big a lexicon is included in the system, it will always be necessary to extend it to meet your own preferences and purposes.

The implemented **grammar** has to be extensive and extensible at the same time. The former means, it is necessary that the system can handle both central and marginal phenomena. The latter means, there has to be a way to pack your own rules into the given grammar, in order to adjust it to the characteristics of your texts. This extensibility is crucial for a pleasant work with the system in the long run.

2.3 Entrepreneurial Criteria

Before buying an expensive MT system for your enterprise, make sure to find the answer for the following questions:

- Can the company I am going to buy the MT system from provide me with all the language pairs that I need or will need in the future?
- Is the system appropriate for the kind of translation work in my enterprise?
- Can the system's components easily be integrated in my present configuration or do I have to change everything so I can use the new system?
- Does my staff have enough know-how to work with such a system, to configure it and to maintain it, or do I have to consider additional expenses for training, extra support or maintenance, etc.?
- What internal changes will the introduction of such a system cause? What are the psychological consequences? Will I be able to motivate my staff? (Cf. the article by Volk and Oberholzer in this volume for more information on the use of MT systems in practice.)

3 Is there a Way to Judge Translation Quality?

Now that we have introduced the technical, linguistic and entrepreneurial criteria, we have to ask how the MT systems can be tested according to these criteria. There are some remarks to be made concerning evaluation methods.

First attempts in the field of MT evaluation (as reported in [WOO94]) were based on some false hypotheses. One of them can described as follows: If you translate a text from Italian into French, for instance, and then back to Italian, you will get an exact copy of the original text — provided that your translation system worked well. And every difference between the original and the translated Italian text will indicate an error your system has made. This hypothesis is not correct — consider that even an extraordinarily good human translation will not reach this goal. Besides, to translate in such a manner is not the aim of translation anyway.

A second hypothesis is based on the same misconception. You choose a text to have it translated by a machine and also by a good human translator, and then you compare the output. You are tempted to say: the closer the results, the better the machine translation. Again, any difference between the two texts counts as an error for the MT system. This view is wrong again, for similar reasons. It is possible to translate a sentence in two different ways and to get two different and nevertheless correct results. The machine translated output can be of the same quality although only similar to the human translation.

Of course, this does not mean that the qualified achievements of human translators have lost their status as a measure for any kind of translation. It is still desirable to reach the human translation level. As soon as we are no longer able to decide whether a text has been translated by a machine or by a person, we have reached our goal. We can compare this to the situation in the *Turing test*. Turing proposed to define the intelligence of a computer according to the following setup: You sit on a computer and enter some questions which are answered immediately. If you cannot decide whether the answers are given by the computer itself or by a person sitting in front of a terminal next door, the system shows intelligent behavior.

One way to evaluate an MT system, which avoids the problems mentioned, is to check whether the target text still contains the information of the original. For this purpose, you may give a target text to persons who do not know the source language and ask them questions about the text. These questions must be chosen in a way that they can only be answered with information from the text. In this way you can gain knowledge of how much of the original information is preserved in the translated text.

Another tricky problem in the evaluation process is the judgement of translated texts by experts. Whatever you do, you will get individual and subjective results. Our tests have shown that it is possible to flatten these discrepancies by asking the participants for a relative ranking of some target texts rather than for an absolute rating.

As reported in the literature (e.g. in [WOO94]), you should not compare different language pairs or different translation directions. Otherwise, you will not be able to interpret your results since you do not have a common ground for a valid comparison.

Other reports ([JDB93], [Fla94]) have shown that it is possible to look at the errors that appear in a translated text and to classify them. You can examine to what extent an error affects the understanding of a text (intelligiblity), or how easy it is to correct the error (improvability). It is more difficult, if sometimes not impossible, to find out why the translation system has made a certain error. Such diagnoses are often purely hypothetical. Nevertheless, error classification seems to be a useful tool for MT evaluation.

4 Conclusion

Our tests have shown that a systematic approach for an evaluation of machine translation (MT) systems is possible. Some criteria are easily tested, e.g. the size of the lexicon or the robustness of the system. You can do this with the help of a questionnaire which organizes the assessment of the various system features. Lexicon and grammar testing is an extensive task; you might use lists of words chosen according to their frequency, and test suites of sentences with different grammatical phenomena. Based on the results, you can establish an error classification.

Other criteria are not so easy to measure and need to be determined with experiments. For these, try to find volunteers who all have about equal knowledge of the languages involved and who also have a similar level of computer skills. In order to get valid results, you have to repeat the experiments with a number of persons. Let them explore the systems, and write a protocol or videotape the sessions for further analysis.

References

[Fla94] Mary A. Flanagan. Error classification for MT evaluation. In *Technology part-*

nerships for crossing the language barrier: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas, pages 65–71, Washington,DC, 1994. Association for Machine Translation in the Americas.

[JDB93] P.W. Jordan, B.J. Dorr, and J.W. Benoit. A first-pass approach for evaluating machine translation systems. *Machine Translation (Special Issue on Evaluation of MT Systems)*, 8(1-2):49–58, 1993.

[Kin96] Margaret King. Evaluating natural language processing systems. *CACM*, 39(1):73–79, 1996.

[WOO94] J.S. White, T. O'Connor, and F. O'Mara. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Technology partnerships for crossing the language barrier: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, pages 193–205, Washington,DC, 1994. Association for Machine Translation in the Americas.