13. Evaluation of Machine Translation Software

Shri. S.Chandra, STQC, DIT, MIT

1.0 Introduction:

The topic of MT is one that we have found sufficiently interesting to spend most of our professional lives investigating. But whatever one may think about its intrinsic interest, it is undoubtedly an important topic — socially, politically, commercially, scientifically, and intellectually or philosophically.

The social or political importance of MT arises from the socio-political importance of translation in communities where more than one language is generally spoken. So translation is necessary for communication for ordinary human interaction, and for gathering the information one needs to play a full part in society. Being allowed to express yourself in your own language, and to receive information that directly affects you in the same medium, seems to be an important, if often violated, right. And it is one that depends on the availability of translation. The problem is that the demand for translation in the modern world far outstrips any possible supply. Part of the problem is that there are too few human translators, and that there is a limit on how far their productivity can be increased without automation. In short, it seems as though automation of translation is a social and political necessity for modern societies, which do not wish to impose a common language on their members. For countries like India, Canada and Switzerland, and organizations like the European Community and the UN, for whom multilingualism is both a basic principle and a fact of every day life.

The commercial importance of MT is a result of related factors. First, translation itself is commercially important: faced with a choice between a product with an instruction manual in English, and one whose manual is written in Japanese, most English speakers will buy the former — and in the case of a repair manual for a piece of manufacturing machinery or the manual for a safety critical system, this is not just a matter of taste. Secondly, translation is expensive. Translation is a highly skilled job, requiring much more than mere knowledge of a number of languages, and in some countries at least, translators' salaries are comparable to other highly trained professionals.

2.0 Evaluation of MT

The evaluation of MT systems is a complex task. This is not only because many different factors are involved, but also because measuring translation performance is itself difficult. The first important step for a potential buyer is to determine the translational needs of the organization. Therefore, the needs to draw up a complete overview of the translational process, in all its different aspects. This involves establishing the size of the translation task, the text type of the material and its form (is it machine readable and if so, according to which standards). It also involves considering organizational issues, e.g. the tasks of each member of staff concerned in some way with translation.

2.1 Engine Performance: Speed

In some circumstances, the speed at which the engine churns out raw translated text won't actually be crucial. If the system requires interaction with the translator whilst it is translating, then of course it should not amble along so slowly as to keep the translator waiting all the time. But if it is functioning without direct interaction, it can proceed at its own pace in the background whilst the translator gets on with other jobs such as post-editing or hand translation of difficult material. This aspect also depends on the user's translational needs: if the user's material requires 15 hours daily on a fast MT system and 20 on a slower one, no one will notice the difference if the system is running overnight. Of course, there are situations where the quick delivery of translation output is essential. But in general, slow speed is the one component of MT performance of which upgrading is relatively easy: by buying some faster hardware for it to run on.

2.2 Engine Performance: Quality

This is a major determinant of success. Current general purpose commercial MT systems cannot translate all texts reliably. Output can sometimes be of very poor quality indeed. It is to be mentioned that the post-editing task (and with it the cost) increases, as translation quality gets poorer. In the worst case, using MT could actually increase translation costs by tying up translators in editing and maintenance tasks, ultimately taking up more time than would have been required to produce translations entirely by hand. Because of its enormous influence on the overall translation cost, translation quality is a major aspect in MT evaluation.

3.0 Test Parameters:

3.1 Intelligibility:

A traditional way of assessing the quality of translation is to assign scores to output sentences. A common aspect to score for is Intelligibility, where the intelligibility of a translated sentence is affected by grammatical errors, mistranslations and untranslated words. Some studies also take style into account, even though it does not really affect the intelligibility of a sentence. Scoring scales reflect top marks for those sentences that look like perfect target language sentences and bottom marks for those that are so badly degraded as to prevent the average translator/evaluator from guessing what a reasonable sentence might be in the context. In between these two extremes, output sentences are assigned higher or lower scores depending on their degree of awfulness—for example, slightly fluffed word order ("... in an interview referred Major to the economic situation..." will probably get a better score than something where mistranslation of words has rendered a sentence almost uninterruptible ("...the peace contract should take off the peace agreement....). Thus scoring for intelligibility reflects directly the quality judgment of the user; the less she understands, the lower the intelligibility score. Therefore it might seem a useful measure of translation quality. Is there any principled way of constructing an intelligibility scoring system? Or rather is there any generally agreed, and well motivated scoring system? We do not know of any. The major MT evaluation studies, which have been, published report on different scoring systems; the number of points on the scoring scales ranging from 2 (intelligible, unintelligible) to 9. The 9-point scale featured in the famous ALPAC Report and was not just used to score the intelligibility of MT, but also of human translation. As a consequence the scale included judgments on fairly subtle differences in e.g. style. This scale is relatively well defined and well tested. Nevertheless we think that it is too fine-grained for MT evaluation and leads to an undesirable dispersion of scoring results. Also, we think that style should not be included because it does not affect the intelligibility of a text. On the other hand, a two point scale does not give us enough information on the seriousness of those errors which affect the intelligibility. (A two point scale would not allow a distinction to be drawn between the examples in the previous paragraph, and complete garbage, (or something completely untranslated) and a fully correct translation.) Perhaps a four-point scale like the one below would be more appropriate.

An Example Intelligibility Scale

1: The sentence is perfectly clear and intelligible. It is grammatical and reads like ordinary text.

2: The sentence is generally clear and intelligible. Despite some inaccuracies or infelicities of the sentence, one can understand (almost) immediately what it means.

3: The general idea of the sentence is intelligible only after considerable study. The sentence contains grammatical errors and/or poor word choices.

4: The sentence is unintelligible. Studying the meaning of the sentence is hopeless; even allowing for context, one feels that guessing would be too unreliable.

3.2 Accuracy

By measuring intelligibility we get only a partial view of translation quality. A highly intelligible output sentence need not be a correct translation of the source sentence. It is important to check whether the meaning of the source language sentence is preserved in the translation. This property is called Accuracy or Fidelity. Scoring for accuracy is normally done in combination with (but after) scoring for intelligibility. As with intelligibility, some sort of scoring scheme for accuracy must be devised. Whilst it might initially seem tempting to just have simple 'Accurate' and 'Inaccurate' labels, this could be somewhat unfair to an MT system, which routinely produces translations, which are only slightly deviant in meaning. The evaluation procedure is fairly similar to the one used for the scoring of intelligibility.

However the scorers obviously have to be able to refer to the source language text (or a high quality translation of it in case they cannot speak the source language), so that they can compare the meaning of input and output sentences. As it happens, in the sort of evaluation considered here, accuracy scores are much less interesting than intelligibility scores. This is because accuracy scores are often closely related to the intelligibility scores; high intelligibility normally means high accuracy.

3.3 Speed

The machine translation software shall be evaluated for the translation speed at word level & sentence level. The total translation time for the text or file shall be measured and speed shall be calculated in character/sec or word/sec.

3.4 Conclusion

The evaluation of machine translation softwares is a complex area. The machine translation software can be for one language or multiple languages and it may have support for one domain e.g. admin or technical etc. or it can have the support for multiple domain e.g. admin, technical, legal, medical etc. For the testing & evaluation of the above-mentioned parameters and for all the languages and domains it is required to have sufficient test data to cover all the functional areas. The user may be more interested for the functions like accuracy, intelligibility or speed etc., but for the test & evaluation of software product, standards like ISO/IEC 9126, ISO/IEC 14598, ISO/IEC 12119 are also available where guidelines for test & evaluation is described in details.