Report of the Session 5 of Tralogy I : Quality in Translation

Joseph J Mariani

Report of the Session 5 of Tralogy I : Quality in Translation

Joseph Mariani

TRALOGY

Report of the Session 5 of Tralogy I : Quality in Translation

Joseph Mariani

LIMSI-CNRS & IMMI, France

Chair: Josef van Genabith (CNGL, Ireland)

Rapporteur: Joseph Mariani (LIMSI-CNRS & IMMI, France)

TRALOGY I - Session 5
Date d'intervention : 04/03/2011

The chair, Josef van Genabith, from the Centre for Next Generation Localization (CNGL), opened the session, mentioning that quality evaluation is a very important topic for translation in general, and especially for Machine Translation (MT). He reminded the audience that this topic was already stressed in the previous session on Translators' tools. He introduced the three speakers, mentioning that they both conduct research on Machine Translation and on MT evaluation, the rapporteur and the discussants.

He then introduced the first speaker, Philipp Koehn from the University of Edinburgh, as the mind behind the Open Source Moses MT system, and his talk entitled "What is a better translation? Reflections on six years of running evaluation campaigns".

P. Koehn expressed that the goals for evaluation metrics should be: to be **correct** (it should rank better systems higher), **consistent** (the repeated use of a metric should give the same results), **low cost** (in terms of time an money), **tunable** (optimize system performance towards metric) and hopefully **meaningful** (easy to understand and interpret). Other evaluation criteria than translation quality may be considered: speed of translation, size of the MT device, easiness of integration into workflow and customization of the user's need. Evaluation is very important in the MT development loop, as it allows assessing the appropriateness of a new scientific idea. This development loop may be conducted several times a day, and must rely on a simple automatic computation of the MT quality. Evaluation metrics may be automatic or manual.

In the first case, the MT quality is computed through a comparison with a human translation reference, or usually, with several human translations, as there may be several different correct translations, even from a single human translator. The BLEU metrics is based on the similarity, or degree of overlapping between the string of words produced by the machine and the human translation reference(s). The METEOR metrics conducts a flexible matching, considering stems, synonyms and paraphrases. He expressed the critique of such automatic metrics: the fact that they ignore the relevance of words, that they operate at a local level, without considering the grammaticality of a complete sentence, that the scores are meaningless (is a MT system achieving a 35% BLEU usable?) and that human translators may get BLEU scores worse than machines. He mentioned that there are evaluation of evaluation metrics, and he shows examples of correlations between automatic metrics and human judgment. This correlation may be computed by using the Pearson's Correlation Coefficient. He also shows evidences of shortcomings. He concluded by saying that automatic metrics are essential for system assessment, that the present ones are not fully suited to rank systems of different types, and that better evaluation metrics is still a challenge.

He then introduced human judgment of the quality of a translation. He gave an example of the regular evaluation campaigns at the ACL Workshop on statistical MT, which gathered in 2010 29 institutions and 153 submitted systems translations. Two metrics are usually used: **adequacy**, giving a measure of the correct translation of the meaning, and **fluency**, giving a measure of the grammatical correctness of the translation, both on a 1-5 scale. He showed examples of evaluators' disagreement on the assessment of the same translation, saying that some evaluators even disagree with themselves (!), and the way to measure the agreement between evaluators. A simpler and more consistent way to express a judgment on a translation is to rank two translations. He mentioned the possibility to give evaluators stricter guidelines, and to adapt penalties to the importance of the translation error, but without much hope. Other approaches are more related to the completion of a specific task, such as providing help to a human translator, or gathering information from multilingual sources. Regarding the first task, a measure of quality is to compare translation from scratch and post-editing the result of an automatic translation. This type of evaluation is however time consuming and depends on the skills of the translator and post-editor. Some metrics are inspired by this task, such at TER, based on the number of

editing steps, or HTER, computing the editing steps between an automatic translation and a reference translation. A final approach is based on understanding tests, such as asking questions to a human reader on the content of the machine translation output. But it's very hard to devise questions. A simpler approach is to evaluate how understandable a translation is in two steps: in the first step, a human edits the translation to make it as fluent as possible (including the possibility to mention that no corrections are needed, or that it's impossible to correct); in the second step, another human indicates whether this edited version is in agreement with the source text and the reference translation. While the inter-evaluator agreement is much better than when using sentence ranking, it however appears that the ratio of edited sentences judged as correct may go down to 69% (reference sentence) and 22% (machine translated sentence), depending on the language pairs. Further findings show that human translators vary significantly in strictness. They are influenced by recently observed translation quality, and they may give professional translations only a 60% score! Crowdsourcing has been used recently to achieve MT evaluation, but it requires the control to check the quality achieved by the contributors.

In conclusion, Philipp Koehn wondered whether the initial goals of MT evaluation have been achieved: **correctness** of evaluation metrics are very hard to determine for manual metrics and measurable for automatic metrics by correlating with human judgment; **consistency** and **low cost** are achievable by automatic metrics, not manual ones; only automatic metrics are **tunable**, and none of the metrics are **meaningful** and easy to explain. There is therefore room for more investigations and further improvement.

Christian Federmann, from DFKI (*Deutsches Forschungszentrum fuer Kuenstliche Intelligenz*), then raised the issue of "How can we measure machine translation quality?", and more generally of "What is a good translation?". He mentioned that MT is a complex task, with various approaches including Rule-Based MT (RBMT), Statistical MT (SMT) and Hybrid MT, which merges various approaches. He expressed that there should be a better bridge between MT researchers and human translators. He also mentioned that MT quality evaluation is itself a complex task, and raised the BLEU metrics dilemma: SMT relies on automated scoring such as BLEU for tuning, and the correlation between such a metrics and human judgment is problematic. Therefore, using this kind of metrics may wrongly shift the problem from "What is a good/acceptable/bad translation?" to "What is the best scoring translation?" (according to a score which is used to tune the MT systems based on statistical approaches, thus introducing a bias). He then gave a brief recap of the automatic metrics and of manual evaluation approaches, also mentioning Phrase-based ranking in addition to Sentence ranking. He thinks that evaluation should be more user-centric, and, while manual evaluation is too costly, that new semi-automatic metrics should be aimed at, taking more semantics into consideration. In conclusion, he proposed to go from automatic to manual evaluation by integrating more human knowledge into MT systems, and to go from research to consumers, taking more into consideration user needs and avoiding "research for researchers only". As mentioned in previous sessions, "Translation is a human activity". MT R&D should therefore be more human-centric and Google Translate is not the (ultimate) solution.

Finally, John Moran, from the Centre for Next Generation Localization (CNGL), presented "Unobtrusive methods for low-cost manual evaluation of machine translation". He placed MT in the framework of its use by human translators where they are exposed to the output of candidate MT systems on a random sentence-by-sentence basis and the evaluation system would report back on which system requires the least post-editing. He briefly introduced post-editing research, and the way to assess gain in productivity. He described an "instrumented" Computer Aided Translation (CAT) tool that he adapted from an existing open-source CAT tool, which provides human translators with a set of possible translations produced by various MT systems and records post-editing effort on the basis of events like keystrokes or deletions. He then expressed that automatic metrics are only quality indicators. Evaluation competitions that provide expensive manual evaluation data only happen annually. A public domain CAT tool that

meets the needs of both MT researchers and translation companies combined with public domain MT systems could provide a channel for gathering large scale MT evaluation data based on post-editing to give more reliable and regular information on the respective quality of MT systems.

Josef van Genabith then introduced the general discussion, stressing that evaluation and quality insurance are definitely a crucial issue in translation. Dominique Durand-Fleisher, a free translator, expressed that her customers would not accept a translation that is only "understandable". She therefore thinks that there is a gap between MT research and actual applications. John Moran answered by citing a SDL MT survey showing that senior managers of major companies are also interested by information produced in foreign languages, such as Forums, that they may be obtained by quick and inexpensive shallow or post-edited MT. Daniel Toudic (*Université Rennes II*) was pleased to see that, at last, MT researchers take into account all the quality insurance tools which have been used for years by translation companies. He finds obvious that the BLEU metrics is not sufficient, as translation doesn't consider only strings of words, but meaning. Philipp Koehn answers saying that it is necessary to have a quick, inexpensive and reliable metrics for developing MT systems, in order to test new ideas and research directions. J. van Genabith then gives the floor to the discussants. Béatrice Vallantin, from the *Agence Nationale de Sécurité Sanitaire de l'Alimentation, de l'Environnement et du Travail*, mentioned that she both produces translation, and uses translations produced by third parties. She agreed that human judgment is problematic. She raised the issues that source texts themselves may be problematic. In the case when there may be several authors, the human translator can bring consistency among them. Also in the case of changes in company names, only the human translators are able to produce the correct translation, thanks to the knowledge they may have of the company history. Edouard Geoffrois, from the DGA/French Ministry of Defense, asked why automatic metrics are meaningless: if it may be true for the BLEU score, he believes that measures such as TER or HTER are more meaningful. C. Federmann answered saying that the meaning of an improvement of 0.05 is difficult to interpret. Ph. Koehn agreed that such measures are more meaningful than BLEU, although it's still difficult to interpret the exact meaning of an improvement of 10%.

This session showed that evaluation might be considered as the key point in Translation in general, and in Machine Translation in particular. If the user needs a quick, easy and inexpensive overview of what has been said in a language that he doesn't master, then he will accept a poor quality that may be produced by raw MT. If he's looking for high quality, then he must go through human translation that may be helped by Computer Aided Translation according to the type of text and to the will of the human translator. Automatic evaluation metrics are mandatory in MT system development, but they are not yet optimal, as they would need to take into account semantics and pragmatics that are difficult to handle by machines. Therefore, they are still an open research topic per se. They rely on comparisons with human translations, that may themselves not be perfect, and the same even for the source texts. More global human evaluation may be used to assess the quality of MT systems in terms of productivity or comfort, but are less fitted to the algorithmic improvement loop.