Tralogy II. Report of the session 5: Assessing Quality in Machine Translation

Joseph J Mariani

Tralogy II. Report of the session 5: Assessing Quality in Machine Translation

Joseph Mariani

TRALOGY

TRALOGY II
Report of the session 5: Assessing Quality in Machine Translation

Joseph Mariani

TRALOGY II - Session 5
Date d'intervention : 18/01/2013

lien video : http://webcast.in2p3.fr/videos-introduction_a_la_session_assessing_quality_in_mt

The chair, Edouard Geoffrois, from the Ministry of Defense/DGA and the French National Research Agency (ANR), opens the session.

Hans Uszkoreit (DFKI, Germany) gave an invited talk on "Translation Quality Metrics for Human and Automatic Translation". He mentions that this "work in progress" is conducted within the QTLaunchPad project supported by the European Commission, and that he wishes to get feedback from the audience on the proposed ideas. The goal of the project is to produce data and tools for evaluating the quality of translation in the target of QT21: Quality Translation Technology for the 21st century. Here, translation quality is taken in the broad sense, not only for (machine-learning based) Machine Translation, and the human translators should be included in the determination of the quality evaluation, as well as the researchers, the users and the technology providers. He mentions that the translation quality could be placed in 3 categories: the good (green), the bad (blue) and the ugly (red), and that the present free online MT systems, such as Google Translate, and evaluation metrics, such as the BLEU measure, only consider the bad and the ugly, while Europe should aim at good quality translation. Therefore using BLEU only would not be sufficient, as a higher quality translation may actually result in a lower BLEU score. He reminds that the quality of translation is two fold: fluency and accuracy, and should respond to the end-user needs (adequacy). He mentions a quote of Alan Melby which summarizes what is translation quality: "*A quality translation demonstrates required accuracy and fluency for the audience and purpose and complies with all other negotiated specifications, taking into account end-user needs*.". He stresses the fact that the measure TQ of the quality of translation should therefore include fluency F, accuracy A and end-user adequacy E, and should consider both the source text S and the target text T. Therefore, he proposes the following measure of quality: $TQ = (A_t\text{-}A_s) + (F_t\text{-}F_s) + (E_t\text{-}E_s)$, that may be superior to 100% given that the translation may even improve the quality of the source text. This measure should be put in relation with the quality requirement QR_t specifications that may vary with the cost, delay and direct availability of the translation. The structure of the Translation Quality assessment measure takes into account the given profile dimensions, derived from ISO/TS-11669, and considers the language aspects (lexicon, orthography, grammar, meaning, style, punctuation) and the document aspects (structure, layout, objects, marking), with various and flexible possible scoring and rating procedures. The final version should be delivered by the end of 2013, after getting feedback from the industrial and scientific community.

François Yvon (LIMSI-CNRS, France) first reminds the audience that there is a strong research activity on machine translation in France, in laboratories such as LIMSI, LIUM or LIG. He then presents the work conducted on the automatic determination of the difficulty of translation for statistical machine translation, stressing that the machine translation quality is not yet sufficient for publishing the resulting translation without human post-editing. He mentions that in some cases, MT and post-editing may reduce the translation time compared with full human translation, by a factor of 30 to 90%, depending on the type of text and language pairs, and that it is therefore important to be able to predict the quality of a translation, given that MT systems presently don't self-assess the difficulty of a translation or the quality of their translation. A campaign on the evaluation of the ability to estimate the quality of a translation has been conducted at WMT 2012 conference on the English-Spanish language pair. The machine estimate of the quality of each translated sentence, on a 1 (ugly) - 5 (good) scale, was compared with an estimate by 3 human experts. All the 11 participating systems, using similar criteria to estimate the quality of a translation (similar length for the source and target sentences, bigram frequency in target sentence, named entity alignment, etc.), got comparable results, which were overall not satisfying. One of the main reasons was the disagreement of the human annotators on the quality of the translation, which means that the machine should both predict the quality and the possible disagreement on the quality. He concludes by stating that it would be more appropriate to predict the usability of a translation, rather than its quality.

Very conveniently, Niko Papula, from the Multilizer company (Finland), then presented a communication on the Benefits and State of the Art of Automatic, Unsupervised Estimation of MT Quality. His company markets a service which assesses the quality of a set of translations

provided by different MT systems or services, either free or paying, and identifies which is the best translated text and whether it's worth post-editing the result, or rather conduct the translation by hand from scratch. He calls the translation enriched with those quality measures "Qualified Translation", to be compared with "Raw translation". He mentions that using qualified translation may increase the translation speed by 11%, on the language pairs that are presently taken into account (English to French, Italian, German, Spanish and Portuguese). He stresses the fact that using qualified MT helps estimating the workload and therefore the price of a translation, or to establish a fair price for a post-editing work. The present challenges are that not all good translations are detected, that the price for qualifying translation is high, given that it uses paying systems or services, and that it still needs more testing. He concludes his talk by inviting translators and post-editors to discuss with him about their needs.

Olivier Hamon (ELDA, France) then introduces the "A to Z of Manually Translated Parallel Corpora for the Evaluation of Machine Translation", and the protocols established through the experience gained by ELDA, DGA and IMMI in the provision of reference data in several evaluation campaigns (TC-STAR, GALE and Quaero) and for various languages (Arabic, Chinese, English, French, German). It represents a large effort, as producing a test corpus of 22 KWords needs about 50 to 60 working days. The production process was improved over the years, and also now includes speech-to-speech translation. The translation team comprises a bilingual translator and a bilingual post-editor, who are established experts and native speakers of the target language. They are given guidelines: being factual (such as no adding or removing of information, no altering of the order of consecutive segments). It is followed by automatic validation (spell checking, format validation), by quality control (validation by experts according to validation guidelines, and resulting in a validation report, on a sample of 5% of the corpus selected at random, with an allowance of 1 penalty point per n words). Difficulties may appear due to disagreements between the translators and the validators. The lessons learned are that early consolidations are a must to help the translators in their task. O. Hamon also stressed that a 100% correct translation does not exist, but that a perfect translation is not needed given that the evaluated MT systems address the bad to ugly quality segment. Dealing with speech data brings even more difficulties. The structure of the sentences are then well beyond the scholarly learnt syntax. The speech transcription needs to be re-segmented by a team of segmenters, and various speech phenomena have to be taken into account (partial words, reiterated words, onomatopoeia, stuttering, mispronounced words, etc.). Due to the transcription errors, which make it sometimes difficult to understand the transcribed data, it appeared that it was very important to also provide the audio data to the translators. The effort for producing evaluation corpora is very costly, but the resulting data is reusable and sharable.

Finally, Rudy Loock (University Lille 3, France) presents a case study on the relationship between corpus-based translation studies (*traductologie* in French), based on the use of parallel and comparable corpora, and translation quality, using the British National Corpus (BNC) for British English and the Frantext corpus for French within the CORTEX (Corpus, Traduction, Exploration) project. It is generally recognized that a translated text shows differences with a source text in the same language. There are two explanations for that: either the fact that there are universals in translation or that the source text interferes with the target text. The study is based on the second hypothesis, and uses the under or over representation has a quality criterion. He studies the use of adverbs in French (ending by *–ment*) for systematically translating adverbs in English (ending by –ly), which is usually considered as a mark of low quality translation. The study of source texts shows that the frequency of adverbs is lower in French then in English. After a first published study on junior translators students, he presents here a study for 17 Master students, translating novels from English to French, for a total of about 200,000 words. The translations were assessed by human experts and placed in 3 categories: good (A), bad (B) and ugly (C). After suppressing a non-French native student who acts differently, the target texts showed amazingly a difference in the adverb frequency in agreement with those 3 categories (respectively 32%, 38% and 41%). However, this is true globally, but not for each translator. And some professional translations in French also reflect a large number of adverbs. Future work

will check if other linguistic phenomena, such as existentials (There is/are – *Il y a*…), use of passive tense, etc.), may also reflect a difference in translation quality, individually or altogether.

Edouard Geoffrois then introduced the general discussion.

Niko Papula was asked if his measure of translation quality was similar to the BLEU score. He answered that it is different, but also ranging from 0 to 100.

Regarding the last talk, it was raised that this is related to a mark of fluency quality in the translation of texts in literature, which may be of less importance in other kinds of texts, where accuracy may be more important. Rudy Loock agreed on this comment. The point was also raised that humans are worse than machines for identifying if a text is a genuine text or a translated one, and that translator students often get comments saying that their translation is good, but that the translated text doesn't sound as produced by a native speaker. François Yvon mentioned that research investigations are conducted on automatically detecting the source language of a translated text.

Finally, François Yvon was asked if there should be a shift from translation technology evaluation to user evaluation. He answered saying that both are necessary. The targeted translation quality may be different if the translated text is to be simply read, or to be published. He believes the metrics must be adapted to the use. Also he thinks that a bad translation may be easy to post-edit into a perfect translation, and the reverse. Evaluation should be addressed in different ways, with different approaches, taking into account what will be the use of the translated texts.