Chunk accuracy: A simple, flexible metric for translation quality

Lars Ahrenberg
Department of Computer and Information Science
Linköping University
E-mail: lars.ahrenberg@liu.se

Many approaches to assessment of translations are based on error counts. These are usually supported by detailed taxonomies that highlight different quality aspects. Even if provided with guidelines the categorization and localization of errors can be quite difficult and time-consuming for a human annotator. Efforts may be wasted on details that are not really relevant for the purpose at hand. For example, a post-editor may be more helped by getting information on the locations of errors than a detailed classification of them.

A framework such as MQM: Multidimensional Quality Metrics (Uszkoreit&Lommel, 2013) is very helpful as a guide to what may be relevant for a given evaluation purpose. There is still a problem of applying criteria, however, once you have a taxonomy. Even if your selection is small, it is still often multi-dimensional, and ambiguities are likely to arise. For example, the distinction between error categories such as Wrong Translation and Missing Word may be clear in principle, but can be hard to make in a concrete case. Also, the question remains how a multi-dimensional selection is used to compare systems. As Williams (2001: 329) puts it: “The problem is this: assuming you can make a fair assessment of each parameter, how do you then generate an overall quality rating for the translation?"

I suggest that these two problems can be at least partly remedied by the following measures: (1) use the simplest possible taxonomies and give priority to counts before types; (2) use chunks as the loci of problems; a chunk can be read as a single unit by a human and eases the task of assigning a problem to a particular word, as for instance in the case of agreement errors. Still, it is more informative than counting errors for the whole text or complete sentences. For example, a post-editor may be shown not just that there are errors in a sentence, but in which part of the sentence the errors are located.

In the simplest case chunks need only ve categorized into problematic (P) and correct (C). The metric then becomes C/(C+P) (or a percentage). To increase granularity, we can use a n-ary scale (for example good, bad, and ugly as is currently popular) and report a distribution over these categories. To get more informative we can categorize problems as those pertaining to adequacy (relation to corresponding source chunks), fluency (target language problems) and others. And then climb further down a taxonomy such as the MQM as motivated by the evaluation purpose.

Chunk accuracy can be applicable whenever a micro-level analysis is called for, e.g., in assessment of student translations, in post-editing settings, or even for MT development. It can be automated to a some extent, thus reducing the human effort. While aggregating observations at the micro-level, and reporting quality characteristics, it is not a final assessment, however. It reports values, not thresholds.

In my presentation, I will further develop my arguments and make comparisons of chunk accuracy to other known frameworks for error analysis.

References

H. Uszkoreit and A. Lommel (2013). Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment.(http://www.qt21.eu/launchpad/sites/default/files/MQM.pdf)

M. Williams (2001). The Application of Argumentation Theory to Translation Quality Assessment. Meta 46(2): 326-344.