Quality Evaluation of Four Translations of a Kidney Document: focus on reliability

Abstract

This paper describes the Kidney project, which began as an experiment to determine whether human translation and fully post-edited machine translation are interchangeable and if so which is more efficient. In the experiment, an English-language patent dealing with kidney cells was translated by a professional human translator and by a commercial machine translation system. The raw machine-translation output was then fully post-edited by three other translators. Thus, four translations of the Kidney patent were available. When the four translations were evaluated by professional human translators, it was found that the evaluation results were not sufficiently consistent with each other. That is, the evaluation process was not sufficiently reliable. The focus of the Kidney project then turned to increasing reliability by analyzing evaluations linguistically to decide how to develop a revised evaluation instruments. As of September 2015 the analysis is in progress. When the revised metric is available, translators not previously involved in the project will be trained and will apply the metric to the same four translations to determine whether reliability has increased or decreased. The Kidney project is being conducted within the MQM framework (http://qt21.eu/mqm-definition), which was developed under the leadership of DFKI (http://www.dfki.de/lt/).

1. Credits

The Kidney project is a collaborative effort of the Translation Research Group at Brigham Young University (Provo, USA), and the Tradumàtica Group at Universitat Autònoma de Barcelona (Bellaterra, Spain). The main participants are Daryl Hague, Pilar Sanchez-Gijon, Kekoa Riggin, Carla Ortiz, and Alan Melby. We thank DFKI for use of MQM.

2. Some Background on the Kidney Project

This paper is an interim report on an on-going project whose focus is to increase the reliability of translation quality evaluation in a particular environment, namely, patent translation for the purpose of filing with a patent office in another country. The project described in this paper is called the Kidney project because it is based on a medical industry patent about kidney cells. However, it is hoped that the results of this project will be applicable to other translation environments, after appropriate adaptation to particular requirements.

Logically, any project involving evaluation of translation quality would begin by defining translation quality, although this is seldom done in practice. The quality of a translation, regardless of how it is produced, can be defined as the degree to which it meets agreed on specifications, so long as those specifications take into account the needs of the intended end users. Of course, some would challenge this definition. Various perspectives on translation quality are presented in issue 12 (December 2014) of the journal of the Tradumàtica group (http://revistes.uab.cat/tradumatica/issue/view/5).

The Kidney project is based on the MQM framework, which has adopted a specifications-based definition of translation quality compatible with the one in the previous paragraph.

The MQM framework is being developed at DFKI (http://www.dfki.de/lt/). See http://qt21.eu/mqm-definition for the official definition of MQM and note that MQM has accepted ASTM International standard F2575-14, Section 8, for defining structured translation specifications (see www.astm.org and search for F2575 to obtain a copy of this standard). For readers familiar with TAUS DQF (see https://evaluate.taus.net), it is relevant that in parallel with the Kidney project, MQM and DQF have been harmonized, under the QT21 project (see http://www.qt21.eu/). Thus, the next stage of the Kidney project will be both MQM and DQF compatible, and when MQM is mentioned, it should be understood as the MQM-DQF approach.

MQM has a broad scope of application. One way to divide up types of translation to be evaluated is by how a translation is produced: classic human translation at one end, raw machine translation at the other end, and post-edited machine translation in the middle. MQM is intended to apply to all three types. The QT21 project emphasizes evaluation of raw machine translation, within a larger context of developing new methods for machine translation. The Kidney project involves human translation and post-edited translation. Thus, the MQM aspect of the Kidney project and the MQM aspect of the QT21 project are complementary.

At this point, it is important to note that the MQM approach to translation quality evaluation contrasts with typical translation quality evaluation methods that use one or more reference translations and an automatic metric such as BLEU. MQM metrics do not use a reference translation but do require the involvement of a professional human evaluator. The homepage of the QT21 project (http://www.qt21.eu/) indicates that along with developing new techniques for machine translation, an important QT21 objective is "improved evaluation and continuous learning from mistakes, guided by a systematic analysis of quality barriers, informed by human translators". It is recognized in the QT21 project and elsewhere, based on widely accepted principles of assessment theory and practice, that reliability is always important and can be difficult to achieve when human evaluation is used.

Those readers familiar with BLEU and other automatic evaluation methods might ask why go back to human evaluation, after it was rejected years ago as too costly and unreliable. (See, for example, "Evaluation of Machine Translation and its Evaluation", Joseph P. Turian, Luke Shen, and I. Dan Melamed, New York University, 2006, Accession Number ADA453509). The motivation in both the Kidney project and the QT21 project for putting humans in the loop is the same: "In order to improve quality, reliable and informative quality measures are required" (QT21 project proposal). The QT21 project proposal goes on as follows: "Although very efficient for quick development of systems and for estimating overall quality, metrics such as BLEU ... are not able to work at different levels of granularity, distinguish between different types of quality problems and give any details about the nature of errors." That is, they are not informative about exactly what to do to improve the system.

The Kidney project team is not claiming that MQM-style evaluation will replace BLEU-style evaluation of raw machine translation. However, we do predict that MQM will become an important factor in evaluation of various types of translation when an informative evaluated is needed, if questions of reliability can be addressed in a satisfactory manner. Thus, the focus of the Kidney project is reliability.

3. Project Description

This section indicates where we are with the Kidney project as of early September 2015. An update will be provided at the MT Summit in late October.

Given a set of translation specifications, the MQM framework can be used to develop a customized translation quality metric. This is exactly what was done in the Kidney project. The Kidney metric is tailored according to the specifications, including the purpose of the translation, which is submission to a patent office in Latin America.

In December 2014, an experiment was conducted to investigate the use of post-edited MT to efficiently produce acceptable translations of patent applications. The experiment aimed to answer the following research question within the larger investigation: Using particular instruments, including a customized MQM metric and specialized training material, can human translators produce a reliable evaluation of the quality of human translation and fully post-edited machine translation? A related research question was whether, in this case, human and post-edited machine translations are indistinguishable on the basis of translation quality evaluation. Any solid conclusions regarding this second question require reliable evaluation and thus an answer to the first question. These two research questions are relevant to a determination of whether post-editing results in acceptable patent translations. Questions of efficiency, while important, are beyond the scope of the current investigation.

In the December 2014 experiment, an English-language patent dealing with kidney cells was translated by a professional human translator and by a commercial machine translation system. The raw machine-translation output was then fully post-edited by three other translators. Thus, four translations of the Kidney patent were available. When the four translations were evaluated by professional human translators who had not been previously involved, it was found that the evaluation results were not sufficiently consistent with each other. That is, the evaluation process was not sufficiently reliable. The focus of the Kidney project then turned to increasing reliability by analyzing the evaluations linguistically and developing a revised metric and associated training material for human evaluators. The question of testing the competence of the evaluators must also be addressed. As of September 2015 the analysis is in progress. When the revised metric is available, translators not previously involved in the project will be trained and will apply the metric to the same four translations, so that we can determine whether reliability has increased or decreased.

In total, seven human translators took part of the December 2014 phase of the project: while four of them participated by translating and post-editing the patent respectively, the other three participated by evaluating the human translation and the fully post-edited machine translation.

4. Discussion of the Results

We are currently in the linguistic analysis phase. The Kidney project team is looking at the first 300 translation units. We have three evaluations of the human translation, and we are examining the differences in how the evaluators annotated each translation unit. In the majority of the translation units, the three evaluators completely agreed. That is, they either all three indicated that there were no errors or all three indicated that there was at least one error and agreed on what the error was.

We are now examining in detail the translation units where there was disagreement among the evaluators. For example, the phrase "prepared from a human kidney-derived cell" appears several times in the patent. There is some debate about the relationships among the constituents and how they affect a translation into Spanish. Is the cell derived from a human kidney or is it a human cell derived from a kidney? Does it make a difference to a patent examiner? Another example is how the linguistic expression "such as" is translated into Spanish in various contexts. Is there any agreement between this expression and other elements of a sentence? A third example is how the word "removed" should be translated into patents in various contexts.

5. Further Work

The Kidney project is far from over. Once we have completed our analysis of the disagreements among the evaluators of both the human and post-edited translations, we will revise the translation quality metric, taking into account the recently completed MQM-DQF harmonization, and improve the training and screening material for evaluators. For one thing, we will give the evaluators access to terminology database. We will probably also develop a tool to help the evaluators deal more efficiently and consistently with multiple instances of the same error. Then we will run the evaluation portion of the December 2014 experiment again, this time with a new set of evaluators who have not yet been involved in the project.

Hopefully, an analysis of the second evaluation of the same four translations will reveal more reliable results and the techniques we use to increase reliability in the Kidney project will apply to other environments where translation quality evaluation needs to be informative or where there not reference translation is available.