Potential impact of QT21

Eleanor Cornelius
Department of Linguistics
University of Johannesburg,
South Africa
International Federation of
Translators (FIT)
eleanorc@uj.ac.za

Abstract

This paper describes the QT21 project from the perspective of the International Federation of Translators (FIT) in three main parts. Firstly, six of the ways that humans currently relate with machine translation (MT) systems will be outlined, leading up to a seventh way that will be discussed in more detail. Huge volumes of texts need to be translated in different sectors of the economy globally. A feasible approach to meeting this need is to employ both raw MT and humans, including translators, in addressing the world's translation needs. Secondly, analytic evaluation of MT quality by human translators will be introduced, focusing on the MQM framework. This seventh way involves annotation, by humans, of specific errors in the raw MT using standardized error categories, rather than only generating a single number indicating overall quality. Lastly, the potential impact of QT21 on MT and professional translators will be reflected on. Through FIT, human translators will be able to participate in the development of improved MT systems. This will help them give objective advice to clients and to guide the developers of next generation translation tools. FIT's position is there will be enough work for translators who do not feel threatened by MT.

1 Introduction

The primary aim of this presentation is to determine the potential impact of the QT21 project on human translators. In order to do so, I firstly provide some background information on the International Federation of Translators (FIT), and on the QT21 project, including FIT's involvement in this project. In the sections that follow, various ways in which humans interact and engage with machine translation are listed and described. A discussion of the QT21 project will not be complete without focusing on two important aspects of this project, namely research and evaluation. Lastly, I consider the impact of the QT21 project on human translators in the years to come.

2 Introduction to FIT and description of the QT21 project

FIT is an international federation of associations of translators, interpreters and terminologists. Through affiliation, more than 80 000 translators in 55 countries across the globe are represented in FIT. In short, FIT's goal is to promote professionalism in the disciplines it represents (http://www.fit-ift.org/).

FIT is a partner in the three-year Quality Translation 21 project (abbreviated as QT21 project) which runs from February 2015 to February 2018. QT21 is a machine translation project which forms part of the EU Horizon 2020 Framework. This project is managed by the German Research Center for Artificial Intelligence (in English abbreviated as DFKI). The main purpose of the QT21 project is to address language barriers in Europe that impede free flow of information. This purpose is in line with the EU's objective, to be achieved by 2020, for a European Digital Single Market that can operate without any barriers, linguistic or otherwise One goal of QT21 is to improve MT models and outcomes for language that (1) are

Proceedings of the 38th Conference Translating and the Computer, pages 10–18,
London, UK, November 17-18, 2016.

morphologically complex, (2) have free and diverse word order, and (3) are under-resourced. (See http://www.qt21.eu/.)

The explosive growth in data witnessed today has not seen an equal growth in the level of translation. MT can go a long way to address this imbalance between supply and demand, notably in cases where its quality is sufficient for the purpose at hand without taking away from the current work of translators. This relates to work not currently done by human translators.

Through investigation and analysis of innovative methods of machine translation, QT21 and FIT will engage translators in assessing the quality of machine translation, incorporating human judgement into the current data-driven development paradigm. Analytic metrics developed in the context of QT21 have already seen the harmonization of MQM and DQF into a single framework, an early deliverable of the project, to define benchmarks for translation quality. Indeed the project proposal points to the need for "metrics [that apply to both] human and machine translation." (See http://www.fit-ift.org/introduction-to-qt21/.) Through contracts between FIT and DFKI, FIT is also instrumental in the dissemination of the findings and advances of the QT21 project.

But how do human translators currently engage with MT?

3 Human engagement in machine translation

In this section, I list and discuss six of the various ways that humans, including translators and non-translators, currently relate with machine translation (MT) systems. At the end of section 4, I describe a seventh way.

3.1 Provision of input

Human translators provide input to MT, in the form of training material for data-driven systems. Pre-processing involves making source texts and their translations into bitexts and includes normalisation of those bitexts. A bitext, according to Harris (cited in Melby, Lommel & Vásquez, 2014: 409), "is a source text and its corresponding target text as they exist *in the mind of a translator*. […] Together, the translation units of the bitext constitute the entire source and target texts 'laminated' to each other."

Specifically, normal pre-processing includes the following: [1]

- *Sentence tokenization (segmentation):* This entails putting each sentence/segment on its own line.
- *Sentence alignment*: Ensuring that source and target sentences are on the same corresponding line numbers, and possibly using empty lines when there is a many-to-one or one-to-many sentence relationship.

At this point, a bitext has been created.

- Depending on the MT system, *removal of formatting annotations*, like italics, bold, hyperlinks, etc.
- *Character normalization*, so that orthographic variations are systematic. Examples include: replacing non-breaking spaces with normal spaces; opening and closing double quotes (“,”) with neutral ones ("); same for single quotes; replacing guillemets/angle quotes (« or »), German-style Anführungszeichen, and east Asian quotation marks with consistent ones (when a language uses multiple forms); normalizing combining characters with precomposed characters; some languages use multiple orthographic

[1] I am hugely indebted to Jon Dehdari from DFKI who provided the information contained in this section.

variants, like the use/non-use of the zero-width non-joiner in Persian, use of Eszett (ß) in Germany and Austria vs.double-s in Switzerland and Liechtenstein;

- *Tokenization*: separating punctuation away from words, so that the following sentence:

I'll take 3.5 of those, please.

becomes

I 'll take 3.5 of those , please .

This is a step that seems easy, but is annoyingly difficult to get right, especially across languages. There are currently more tokenization algorithms than there are pubs in Ireland!

- *Lowercasing or truecasing*: Truecasing is making a word lowercase if it normally is. For example, a truecaser would change the first word in an English sentence to lowercase for a word like "The", but not for a word like "Japan". Some words are tricky, like "University" or "Apple".

Data-driven MT heavily relies on human translation. In order words, the human translates the source text that is used as training material in MT. It thus follows that without human translation, there would be no data-driven MT.

3.2 Pre-editing of sources texts

Whereas pre-processing, as discussed above, involves creating bitexts from already translated source texts, pre-editing (PE) involves preparing source texts, that are not part of the training material, for MT. According to Martinez (2003: 16) the aim of pre-editing is "to achieve better human readability and clarity of the SL text, as well as better computational processing or translatability, especially by translation systems" (original emphasis). The distinction between "readability" and "translatability" is discussed further below.

The concept 'Controlled Language' (CL) during authoring is relevant to pre-editing as, according to O'Brien (2010: 143), there is overwhelming evidence that application of CL rules (that is, "constraints on lexicon, grammar and style with the objective of improving text translatability, comprehensibility, readability and usability") has a marked positive effect on MT quality.

Reuther (2003: 124-5) explains that CL can have two uses: (1) to enhance readability and understanding (i.e. cognition) by focusing on text linguistic aspects, and (2) to improve translatability by MT systems. In either case, the processing system may be human (a human reader, or a human translator) or it can be automatic (a monolingual automated language processing system, or an automated translation system, such as TMs or MT systems). Reuther (2003: 125) provides examples of constructions on lexical level, formatting level, and phrase and sentence level, that may pose problems for MT systems and, in some cases even humans as well, to process, regardless of use (to enhance readability and understanding, or to improve translatability). On lexical level, spelling, morphological and synonym variants may create processing problems. On formatting level, issues relating to punctuation, spacing and typography may pose problems for MT systems, but not for human processing. On phrase and sentence level, Reuther (2003: 126) indicates that some syntactical constructions affect readability and comprehension, but do not pose translation problems (regardless of whether the translation is done by a human or an MT system). In other cases, both comprehension and translatability may suffer. Readability CL rules and translatability CL rules are not vastly different, as readability rules are subsumed under translatability rules. This means translatability, according to Reuther (2003), facilitates readability, as there are only a few translatability rules that are not also readability rules (that is, at least, in the case of German).

Even with the use of TM and the resulting quality of the translated output, it is important to feed the TM with controlled output from the very beginning. If this is not done from the outset, a mismatch will occur between controlled input and uncontrolled reference material stored in the TM (Reuther 2003: 131). As (1) texts are written by humans, and (2) CL on all levels (lexical, formatting, and phrase and sentence level, and possibly also global text level) ensures both readability and translatability, the role of the human in the pre-editing process should not be under-estimated.

3.3 Gisting and triage

Humans also perform gisting and triage; that is, they assign a general meaning to raw MT output (gisting) and decide whether further processing is needed (triage). As early as 1979, Henisz-Dostert (1979: 153) cited Garvin (in Lehmann and Stachowitz 1971: 114) who said that MT output, for various purposes, "will be only casually scanned rather than carefully read". Although this source (Henisz-Dostert, 1979) is particularly old, it is interesting to note that not much has changed since that time (as far as gisting is concerned, at least).

This idea of scanning a text, translated by an MT system, is now also known as content gisting or browsing (see Martinez, 2003: 18). Gisting is a monolingual human activity in which the source text has no place or importance (this means the source text is not consulted during the gisting process). The purpose of gisting is to arrive at a general idea of what is conveyed in a translated text, i.e. the raw MT output. The following response of a respondent in Henisz-Dostert's study (1979), to the question "Why do you use MT?" sums up the purpose of gisting particularly well: "To determine if the publication contains material that is pertinent to my work." In such cases, the "speed of access could compensate for inadequacies of machine translation" (Henisz-Dostert, 1979: 155), as MT is faster than human translation (ibid, 166; 184). The person who does the gisting does not have to be a translator, nor does s/he have to be proficient in the source language. Gisting can be done for a number of personal reasons – and indeed Henisz-Dostert (ibid, 180) predicts that "under the conditions of a regular, rigorous service, requests for translations for scanning purposes only would become routine", and/or it can be a step leading to triage.

Triage is used by people who are not translators to determine whether human translation of a machine-translated text is warranted. Triage is thus not a form of translation, but much rather a decision-making process aimed at determining the best way to proceed in order to reach a particular goal. In relation to optimal use of resources, Melby (2016) makes the following statement: "[…] documents that are useful as raw machine translation or are never consulted do not use up valuable human resources for further post-editing or translation." Of course, in instances of incomprehensibility of a raw machine-translated text, the need for post-editing or improvement arises. This could entail requesting retranslation by a human translator, for instance.

Against this backdrop it is important for professional translators, having specialised knowledge and specific expertise, to advise their clients on whether MT is the best option or whether another approach would be better suited to fulfil the particular translation need.

3.4 Post-editing (PE) of MT output

Another way in which humans are involved in MT relates to PE of raw MT output: the correcting of mistakes in the raw machine-translated text.

According to Martinez (2003: 18), inbound translation to understand (assimilation) is not accompanied by PE (in the case of content gisting) or it is supplemented by rapid PE (RPE) (or light post-editing) in order to correct only the most serious errors in order to improve comprehensibility and accuracy. Texts edited in this way usually have a short life-span.

However, outbound translations to communicate (dissemination) require either (1) minimal post-editing (MPE) – in cases of technical texts, such as a set of instructions, with a longer life-span, in cases where cohesion is not all important, or (2) full post-editing (FPE), in cases where high-quality translation is required, or (3) only 10-20% of a document will be edited in cases where 80-90% accuracy is achieved, for instance the fully automated translation of weather reports.

PE, according to Martinez (2003: 20-2), is done by either by translators, revisers, non-linguists (technical experts) or trained specialists in the company, but the type of PE required will also be a determining factor. Martinez (2003: 22) explains: "(T)his new role where efficiency is a priority, could be successfully fulfilled by 'anyone' with ((very) good) bilingual and linguistic skills, involved in the field of communication of information". However, Martinez (2003: 21) also warns as follows when translators are used as post-editors:

> PE is completely different from translating and requires a different attitude to text production as well as certain "ideal" abilities. Sometimes, when MT software offers low quality, translators can become resentful of the fact that they could have produced a better translation from scratch. In most cases, translators find machine-translated texts irritating and rarely enjoy correcting bad translations.

Important in all of this, is the primary instruction to the human. Is the human instructed to translate or to edit? If the human is instructed to translate, the activity is human translation. If the human is instructed to edit machine-translated text, the activity is PE. Thus, PE is not human translation. The amount of time and cost expended, during PE, to achieve a product of high quality should be carefully considered. If excessive effort is required, then human translation – from scratch – should be advised.

3.5 Use of selected segments of raw MT

Human translation, typically, begins with a source text, accompanied by a set of instructions that can either be implicit or explicit. The end result of this activity is a target text. Humans can optimally use various resources while translating. If the instructions are appropriate and if the translated product meets these instructions, the product will be of high quality.

The professional translator consults various resources during the translation process. This typically includes terminology and translation-memory lookup. The translator, however, is free to either use or ignore suggested (real or fuzzy) matches. Likewise, when segments of MT are available, the translator is free to use them or ignore them.

3.6 No use of MT

Humans can also bypass MT; that is, they can translate from the source text using either no translation-specific tools at all or only terminology lookup and/or translation memory lookup, without consulting raw MT output.

3.7 Translators, bilinguals, and monolinguals

Some of the six translation-related human activities described above require the skills of a professional translator, some only require knowledge of both the source and target languages, while others can be performed by monolinguals. With the huge volume of texts that need to be translated in different sectors of the economy in the world today, the only feasible approach meeting this need is to employ both raw MT and humans, including translators, in addressing the world's translation needs. To this end, collaboration between professional translators and the buyers of translation is all important. FIT does not view MT as a threat to professional translators.

4 An introduction to two aspects of the QT21 project

4.1 Basic research

There are ten well-established universities that are partners in the QT21 project (see http://www.qt21.eu/consortium/). Some of the new approaches to MT that will be tried out are RNNs (Recurrent Neural Networks), novel syntactic and semantic translation models, and APE (automatic post-editing). FIT will not be involved in the basic-research aspect of QT21. These new approaches are mentioned only because they all require evaluation of the raw MT output to determine whether they are better than current approaches.

4.2 Analytic evaluation of MT quality by professional human translators

This section introduces the topic of analytic evaluation of MT quality by human translators, as a goal of the QT21 project is "improved evaluation and continuous learning from mistakes, guided by a systematic analysis of quality barriers, informed by human translators" (http://www.qt21.eu/) (own emphasis). This, then, represents the seventh way in which humans engage with MT. It involves annotation, by humans, of specific errors in the raw MT using standardized error categories, rather than only generating a single number indicating overall quality.

The focus here is on the MQM framework as a complement to, not a replacement for, reference-based translation evaluation methods, such as BLEU, that is widely used (Lommel, 2016: 63).

The most common approach to evaluation of an MT system under development is to select a source text and have it translated by a professional human translator. The raw MT output is then automatically compared with the human translation (the reference translation). Changes are made to the MT system, and the same source text is translated again and automatically compared with the reference translation in order to determine whether the change in the system made the output look closer or further away from the reference translation.

What then are the characteristics of translation quality metrics? A system can either be:

- holistic (focusing on the entire text) or analytic (focusing on specific portions of the text)
- reference-based (it requires a reference/sample/model of translation, previously done) or reference-free (no previously translated text is required)
- automatic, and thus fast, or manual, and thus slower.

Additionally, metrics can differ in terms of their validity. In relation to validity, the following question arises: "Does it measure what it is supposed to measure?" Lommel is critical of the validity of reference-based methods. A reference-based metric (such as BLEU) works on the underlying assumption that a particular reference translation is –

> […] a valid measure of quality and the tests designed to demonstrate that validity bias the results because they use a similar method with human evaluators who cannot independently evaluate the translations without the references that are under consideration (Lommel, 2016: 64).

Although BLEU is designed to cope with more than one reference translation, BLEU scores are typically measured by using only a single reference (Lommel, 2016: 64). Moreover, claims that BLEU matches human judgment may also be flawed, as it is not clear what these judgments are about. It is also debatable, according to Lommel (2016: 64), whether referenced-based methods indeed measure translation quality.

Metrics can also differ in terms of degree of *reliability*, begging the question: "Will the metric perform consistently when used by different evaluators during an actual application?" In terms of reliability, BLEU is reliable. Lommel (2016: 64) states the following:

> Because it is mechanical, for a given set of references and a hypothesis BLEU will always generate the exact same score. When the hypothesis changes the score will perfectly reflect the differences. BLEU does not depend on the judgment of an annotator.

Despite the reliability of BLEU as described in the quote above, MT engines are inherently inconsistent (Lommel, 2016: 65), as they may do very well with one part of text, but perform less well in another part.

In the QT21 project an additional method of evaluating the output of an MT system is used. Professional human translators apply a quality metric developed within the MQM (Multidimensional Quality Metrics) framework. The MQM metric will not be automatic but will be analytic; that is, specific errors in the raw MT are annotated by humans using standardized error categories, rather than only generating a single number (such as a BLEU score) indicating overall quality. Thus, the MQM-based metric in QT21 is manual (not automatic), analytic, and highly informative. Additionally, it does not require a reference translation as, in a typical production environment, there is no reference translation available. Why would there then be a need for another translation? Thus, automatic evaluation only makes sense in a MT development context, where a reference is used an evaluation tool.

Based on Lommel's (2016) assessment, it follows that reference-based methods will not always indicate whether the modified translation is better or worse than a previous translation, and for this very reason, it is therefore not as useful as it may seem.

In light of the above, improvements in quality must be meaningful in human terms. It is therefore important to incorporate judgements of human translators in translation quality evaluation. Both types of metrics (automatic and analytic) have a role to play in the assessment of translation quality. However, the strong and weak points of each system should be carefully weighed up. Whereas automatic metrics are fast and good for research and development, analytic metrics provides insight into specific problems and they can discriminate based on differing specifications (or instructions). The single score an automatic system allocates is not meaningful in human terms as it provides little insight into the problems in the translated product and the types of improvement required to enhance quality. Then again, analytic metrics (such as MQM), are slow and more expensive than automatic approaches, and they cannot be used for rapid development. Therefore, both BLEU-style and MQM-style metrics are needed.

MQM is a flexible system for defining metrics (either analytic or holistic), that allows for various specifications. Each general set of specifications will have its own metric (which may be identical to the metric for another set of specifications in some cases). MQM can be used to assess conformance to specifications for each type of translation:

- Raw MT: Does the translation output meet requirements for end-user usage?
- Triage is a downstream use, but we need to know if the translation is good enough for that use.
- PE: Is the translation fluent and accurate enough to support efficient PE? Does the human contribution bring the translation in line with its specifications?
- MT as an option and "classic" human translation: We can evaluate the text for its intended final use.

In light of the above, it is important to note that there can be no single set of specifications that applies to *all* translation. Quality depends on purpose, needs, and scenario. It is possible

to have a variety of measures of quality; however, not all measures will be appropriate for any given translation project. The metrics that are applied to assess translation quality should be in line with the particular specifications (instructions) relating to the translation project. Different metrics give *different* quality scores for the *same* text depending on the specifications, and thus: what is a good translation for one purpose may not be good for another.

For example: Consider a source text that is written in a very high and difficult register, but the text is being translated for use in educating twelve-year-old students. A metric that values absolute fidelity to the source will give a translation that meets specifications a bad score. A different metric that does not penalize changes in register will give a more appropriate score. Thus, changing what is measured produces a new metric.

MQM defines a family of metrics, as no single metric can ever apply to all translation projects.

Why is MQM good for professional translators? This metric provides a way to specify how translators will be judged that *respects* their ability to produce appropriate translations and their right to refuse inappropriate work. The metric is fair, as the criteria that are used for evaluation of quality is made available in advance. Moreover, MQM allows for direct comparison of different methods of translation and reproducible methods of assessing whether a translation meets the mutually agreed upon translation specifications. Lastly, MQM helps translators to understand the strengths and weaknesses of MT.

5. Potential impact of QT21 on MAT and on professional translators

FIT will invite human translators to participate in the QT21 project, from its substantial pool of translators that it represents through member associations. This will provide an opportunity for those translators to gain an insider view of the world of MT and thus better understand its current status. FIT is of the opinion, as stated in its Position Paper on MT (http://www.fit-ift.org/wp-content/uploads/2016/09/MT_pospaper_exit2.pdf), that "(T)ranslators should seek to respond to the new developments in good time and see how to derive benefits for themselves." Through their involvement and active participation in the QT21 project, translators will be able to see the strengths and weaknesses of MT, because reports of FIT's experience with evaluation will be disseminated to the entire FIT community. All of this, in turn, will help human translators give objective advice to those who need translation services and guide those who develop the next generation of translation tools. MT developers will look for ways to improve MT based on the annotations of human translators.

The position of FIT is that there will be more than enough well-paid work in the foreseeable future for translators who do not feel threatened by MT and who can advise others on a team that can use all seven ways of interacting with MT. In an evolving translation market, the volumes of translation work are increasing. This means the pie becomes bigger and bigger, and so the slices of the pie also grow proportionally in size. FIT's position, as expounded in its Position Paper on Machine Translation (http://www.fit-ift.org/wp-content/uploads/2016/09/MT_pospaper_exit2.pdf) is that there will be instances where raw MT output is completely acceptable. In such instances the user of a text simply wants to extract the gist of a text in its basic form (see the discussion of gisting in part 3.3 above). In other instances, there may be highly adverse consequences to raw MT output, for instance when businesses make available unedited MT texts to accompany their products. Such unedited machine-translated texts could damage the corporate image of the company and there could even be product liability implications.

In balancing the huge advances that are made in the field of MT, there can be little doubt that it is in the best interests of the translator community to actively engage with the entire translation industry on MT, in general, and the evaluation of translation quality, in particular.

Translators should become familiar with FIT's involvement in MQM and should acknowledge that both BLEU-style as well as analytic metrics have a role to play in quality evaluation. Those working in the field of MT people are most probably very familiar with BLEU, but may be less knowledgeable about MQM. Through its involvement in the QT21 project and the development of MQM, FIT plays an active role the translation industry.

Acknowledgements

The project QT21 leading to the above results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 645452.

References

Acrolinx. http://www.acrolinx.com/ Accessed: September 25, 2016.

Dehdari, Jon. DFKI. Electronic communication on 22 September 2016.

International Federation of Translators (FIT). 2016. Welcome to FIT. http://www.fit-ift.org/ Accessed: September 16, 2016.

International Federation of Translators (FIT). 2016. Introduction to QT21. http://www.fit-ift.org/introduction-to-qt21/ Accessed: September 14, 2016.

International Federation of Translators (FIT). 2016. Position Paper on Machine Translation. Available at: http://www.fit-ift.org/wp-content/uploads/2016/09/MT_pospaper_exit2.pdf Accessed: September 22, 2016.

Lommel, Arle. 2016. Blues for BLEU: Reconsidering the Validity of Reference-Based MT Evaluation. In Rehm, Georg & Burchardt, Aljoscha (eds). *Proceedings of the LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem*", pp. 63-70.

Martinez, Lorena G. 2003. *Human Translation versus Machine Translation and Full Post-Editing of Raw Machine Translation Output*. Unpublished master's dissertation. Dublin City University.

Melby, Alan. 2016. Interaction between human translation and machine translation (MT) in translation activities. Available at: https://www.fbcinc.com/e/LEARN/e/Translation/presentations/Wednesday/MT-HT-Spectrum-OneSheet_V4b.pdf Accessed September 30, 2016.

Melby, Alan K., Lommel, Arle & Vásquez, Lucía M. 2014. Bitext. In: Chan, Sin-wai. *The Routledge Encyclopedia of Translation Technology*. Chapter 25, pp. 409-424.

O'Brien, Sharon. 2010. Controlled Language and Readability. In: Shreve, Gregory, M. & Angelone, Erik (eds). *Translation and Cognition*. American Translators Association Scholarly Monograph Series XV. John Benjamins Publishing Company, pp. 143-165.

QT21: Quality Translation 21. Available at: http://www.qt21.eu/. Accessed: September 25, 2016.

Reuther, Ursula. 2003. Two in One – Can It Work? Readability and Translatability by means of Controlled Language. In *Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop (CLAW 2003)*, 15th-17th May, Dublin City University, Ireland, pp. 124-132.