Technology Landscape for Quality Evaluation: Combining the Needs of Research and Industry

Kim Harris[1,2], **Aljoscha Burchardt**[2], **Georg Rehm**[2], **Lucia Specia**[3]
[1]text&form GmbH (Berlin), [2]Deutsches Forschungszentrum für Künstliche Intelligenz (Berlin), [3]University of Sheffield

Abstract

Translation quality evaluation (QE) has gained significant uptake in recent years, in particular in light of increased demand for automated translation workflows and machine translation. Despite the need for innovative and forward-looking quality evaluation solutions, the technology landscape remains highly fragmented and the two major consituencies in need of collaborative and ground-breaking technology are still very divided. This paper will demonstrate that closer cooperation between users of QE technology in research and industry to create a holistic but highly adaptable environment for all aspects of the translation improvement process, most signficantly quality evaluation, can lead the way to novel and ground-breaking achievements in accelerated improvement in machine translation results.
Keywords: Machine Translation, Evaluation, Human Translation

1. Introduction

Currently, the approaches and tools applied by research and industry to evaluate the quality of translation differ widely from each other, in terms of both methodology and implementation. Yet, the needs of both consituencies are largely identical: Both want to determine overall translation quality for various purposes, both want to understand the underlying issues – or errors – and fix them, and, most importantly, both want to improve translation output, i. e., prevent those issues from recurring in the future.

While most language service providers primarily perform QE on translations carried out by professional translators, there is a positive trend towards the integration of machine translation (MT) solutions in "traditional" translation workflows (Autodesk, 2011). Consequently, the demand for efficient QE processes to improve the content as it moves through the typical translation cycle has increased. In a 2013 survey performed by the QTLaunchpad Consortium (Doherty et al., 2013) two-thirds of all language industry respondents said they were currently using or planned to use machine translation in their translation business, and almost 70% said they use human evaluation methods to assess the quality of MT output, with only 22% using automatic evaluation metrics such as BLEU and TER.

Language service providers are often bound by the (human) translation technology dictated by their customers or that offers features that make the translation process more efficient and thus more widely accepted by the translator community. A number of LSPs have incorporated MT generated content into these translation environments and succeeded in integrating quality *estimation* tools in their workflows to filter out the automatic translations that are not worth editing. However, these processes neither fully integrate research approaches nor do they directly support the **improvement** of the generated content for future use. MT is still widely seen as a black box, and very few have the resources to invest in closer ties to the research community, in the rare cases where this is actively pursued.

This approach is the complete reverse of that applied by the research community to evaluate MT output. Historically, research has largely relied on automatic evaluation metrics such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and METEOR (Banerjee and Lavie, 2005) to assess the quality of the MT system output for a specific language pair based on one or more reference (human) translations. While the generated score gives an indication of the *overall* quality, it does not provide information on the *reason* for the outcome, nor does it reveal how to improve the translation in the future.

Some research has seen a shift towards human evaluation in the form of MT translation ranking and other primarily non-linguistic evaluations performed to a significant extent by untested and unqualified crowdsourced resources (Graham et al., 2016) or researchers with no translation background (Bojar et al., 2015). The integration of professional translators in the evaluation process is still lagging, largely due to the lack of collaboration with the language industry on a broader scale.

This gap between the these two consituent drivers of machine translation has become somewhat of a conundrum: Commercial LSPs are unable – even unwilling – to invest in their own systems because they have no access to the necessary expertise, no financial resources and see relative stagnation in MT innovation and therefore no business case for the investment. The research community has been sufficiently successful in proving its own results for its own purposes with automatic scoring and minimal human ranking efforts, and therefore sees little reason to invest financially and otherwise in the integration of professional translators into the research loop to find more novel and less automatic ways of looking deeper into the crystal ball.

2. Fragmentation in the Translation Industry

As a result, there is little overlap in the methods and tools currently used by these two groups for quality evaluation and even less interaction between or influence of one over the other in a move towards more interconstituent standardization. This, however, does not only lie in the lack of cooperation between the research community and the language industry, but also in the inherent fragmentation of the processes and tools implemented by either constituency respectively.

2.1. The Question of Quality

The greatest challenge and root of much debate and discord relates to defining quality. What is it exactly? According to (Koby et al., 2014) "*quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs.*" While there are many other scientific definitions similar to this one, in reality quality is whatever the customer wants it to be. This in itself demonstrates just how diverse and heterogenous quality standards and all aspects of translation quality must be and have always been. As a result, the evaluation of this quality poses a significant challenge if the number of factors affecting quality is multiplied by the number of criteria used to evaluate it.

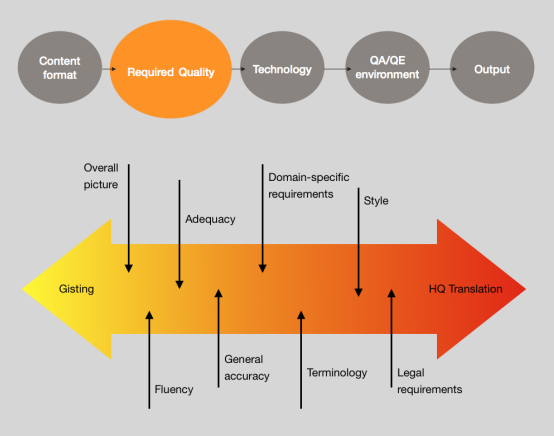

Figure 1: Quality scale in translation workflow

As shown in Figure 1, there are nuances in quality expectations that vary from case to case, and these will depend on a range of factors that infuence the expected quality, including the purpose of the content, its format, domain, time constraints, financial issues and other customer and content-related factors such as tools and publication (Zaret, 2016). All of these factors impact not only the type of quality evaluation performed on the translation but the environment in which evaluation can, or even must, be performed. We can demonstrate the diversity of fit-for-purpose quality expectations by comparing two vastly different scenarios. Customer A requires the translation of a legally binding financial document for broader publication and application. Customer B has general e-mail correspondence between two subsidiaries for internal use only. Not only is the domain different, but so is the purpose. Whereas gisting and some light post-editing might be feasible for Customer B, an absolutely flawless and highly accurate translation will be required by Customer A. Quality for Customer B is proper conveyance of the overall meaning, which is insufficient for legally binding documents.

2.2. Translation Technology Landscape

Given the sheer size, diversity and unabating growth of the language industry, and the lack of standardization in key areas such as format and quality, it is hardly surprising that the industry is enormously fragmented. Translation has become somewhat ubiquitous with the rise of free online translation services such as Google and Bing. Yet, there are over 25,000 registered language service providers worldwide using hundreds of different technologies to perform translation and quality assessment. Fragmentation appears to meet the needs of those who have a demand.

The drive to reach global markets in a competitive landscape has been quintessential in the positive impetus experienced by the language industry, but it has also played a major role in the development of highly specialized and often customized technologies and environments specific to both customer and content. Repositories for open source tools and language resources such as META-SHARE[1] and language technology associations such as LT-Innovate[2] reference hundreds of language tools and resources and demonstrate clearly how significant and how fragmented the language industry is, from both an industry and a research perspective.

The user-driven sophistication of standard technology used by language service providers is striking when compared to that of many open source solutions, particularly those used by the research community. The most successful translation environments are those that offer efficient workflows and features that are profitable to the supplier and provide the level of quality, speed and price required by the buyer of language services. Tools that are too cumbersome or do not support the most common file formats and markup will find little uptake in the industry. SDL Trados Studio™, shown in Figure 2, is currently the most widely used environment for professional translation and MT integration, however, other applications such as MemSource and MemoQ and hundreds of smaller, specialized applications, all of which offer optimized translation features, multiformat support and MT integration and services are on the rise. Needless to say that most tools used by the language industry are neither interoperable nor compatible except in their most basic text form.

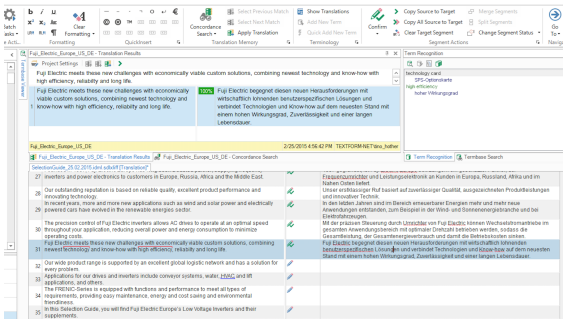

Figure 2: SDL Trados Studio™ user interface

2.3. Quality Evaluation in the Language Industry

While translation technology has experienced a relative boom during the past few years, not least due to the dawn

[1] http://www.meta-share.eu
[2] http://www.lt-innovate.org

of accessible machine translation and the *need for speed* on global markets, standardized, integratable tools to help assess and improve the quality of translated content have not.

The evaluation of translation quality represents an area where the absence of reliable and meaningful standardisation and evaluation methods for buyers, suppliers, MT adopters, among others, is particularly serious (Doherty et al., 2013). As shown in Figure 3, language service providers use a vast number of different evaluation methods and standards to assess the quality of their translation output, with proprietary tools and those integrated in other tools making up two-thirds. This is a clear indication that currently none of the aforementioned translation technologies offer suitable or satisfactory intergrated QE features at the level required by the user, particularly in light of the fact that well over two-thirds of all respondents still use human quality evaluation *only*.

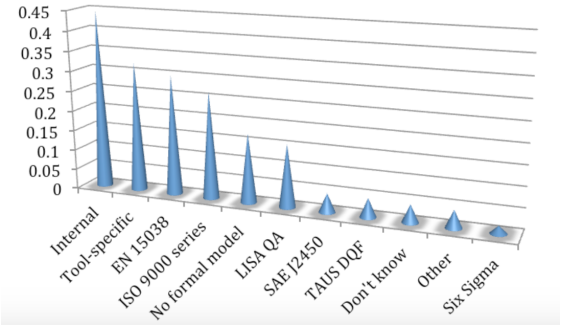

Figure 3: QE tools used by language industry (Doherty et al., 2013)

MT adopters in the language industry do use some of the metrics made available through the research community, such as BLEU and (H)TER, to evaluate the quality of the output, but this is primarily performed in order to filter out what these metrics would consider bad translations based on their scores so that post-editors do not need to do this themselves. It is still considered the most efficient way to perform an overall assessment, but there are no other efficient ways to evaluate translation quality in detail than to do this manually.

MQM, the quality metric developed by the QTLaunchpad Consortium[3] addresses some of these standardization issues with respect to error categorization and the flexible creation of error typologies. It can be integrated into the methods and standards shown in Figure 3, and adapted to fulfill all quality specifications of any given translation task flexibly and easily. This methodology has received positive feedback from a number of research and industry users and has been harmonized with the TAUS DQF [4] to promote industry-wide uptake and push consolidation in the area of quality evaluation.

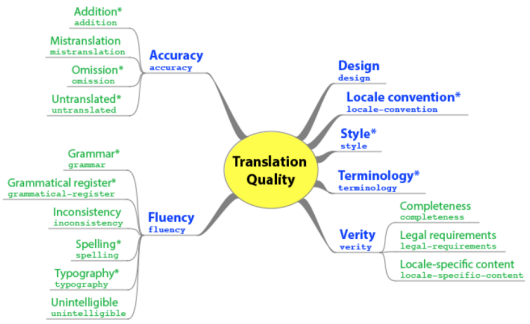

Figure 4: Example of MQM error typology

3. Fragmentation in the Research Community

The research community has been pivotal in the continued development and success of machine translation technologies; however, it is a community that spends much of its effort working in silos, developing tools and solutions for highly specific problems or challenges in a particular aspect of their own research. Rarely does one see a collaborative **interoperable** platform of complementary tools that have the potential to address a larger complex of problems, and even more seldom is an ongoing collaborative effort with the user community of these tools to find applications for them in real-world scenarios. The META-SHARE repository alone boasts 2,725 language resources at the time of writing, 102 of which contain the keyword *quality*.

Automatic evaluation metrics such as BLEU and TER are two of the most popular and inexpensive automated metrics and have been know to demonstrate a relatively high correlation with human judgements. The resulting quality scores are based on comparisons with sets of HT references, which can be useful for certain estimation tasks; however, they do not provide the ability to assess why scores improve or worsen, and they focus almost exclusively on the score, offering insuffucient insight into real error analysis and improvement.

The number of automatic evaluation metrics alone is a clear indication of just how granular an evaluation metric is to a particular subtask of a specific task. Much like the fragmentation found in industry, many of these metrics will have some degree of overlap, yet there seems to be little interest in adapting or combining existing tools instead of developing new ones.

A number of quality estimation and evaluation tools developed by the research community have attempted to combine various aspects of the actual translation quality with the use of automatic metrics, such as QuEst[5], Asiya (see Figure 5) and Appraise[6], the latter of which also integrates human error annotation in its quality metrics.

[3] http://www.qt21.eu/launchpad/
[4] http://www.taus.net
[5] http://www.quest.dcs.shef.ac.uk
[6] https://github.com/cfedermann/Appraise

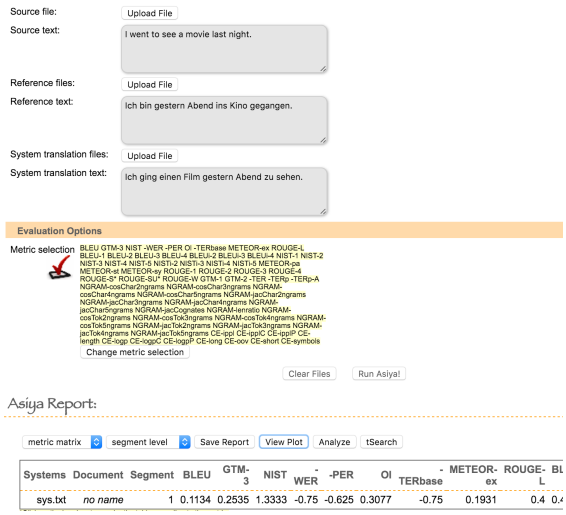

Figure 5: Asiya-Online with evaluation metrics

4. The Motivational Divide Between Research and Industry

As discussed earlier in the paper, the objectives of both communities are identical: to determine overall translation quality for various purposes, to understand the underlying issues – or errors – and fix them, and, most importantly, to improve machine translation output, i. e., prevent occuring issues from recurring in the future. Why, then, have we not seen more cooperation towards these common goals?

Although the objectives are seemingly similar, the motivation that drives them is completely different. Industry, on the one hand, needs reliable, faster solutions that are scalable and financially viable. Quality is no longer a unique selling point. It is a requirement, regardless of how the customer defines it. Settings up machine translation systems and automated quality metrics can be expensive, complex, complicated and embody the proverbial black box for many language service providers. The systems either rely too heavily on large amounts of data and experienced resources with the right background in computer science, or on intrinsic linguistic programming that is time-consuming and only applicable to a handful of language pairs. Neither scenario has proven promising to the majority of LSPs. Real progress is slow, innovative technology drives are few and far between, and the cost of ramping up an MT workflow for a customer often brings with it a significant financial risk.

What is lacking in the language industry is the motivation to participate in a collaborative paradigm shift towards human-informed MT development. There is little interest in collaboration, which stems largely from its cottage-industry heritage, as well as a fear of promoting their own professional demise. Diversity of language is a welcome excuse to remain as fragmented as possible. It is the Darwinian survival of the fittest.

This concept of survival is not unkown to the research community either, and it the force that drives the lone-ranger mentality in many aspects of its work. Most institutions are not interested in finding industry applications for their research but choose to focus on proving the point of their research in order to find and receive funding.

As with language service providers, financial considerations are the key factor when deciding how to spend a budget. Working with industry partners is understandably more expensive than hiring primarily unqualified Mechanical Turkers or finding colleagues or crowd-sourced resources to perform some of the manual tasks involved in some research. It is little wonder that the results are far from ideal, although research would be hard-pressed to agree that lack of skills and qualification may be the cause, but the investment in much more promising collaboration with professionals is seen as too time-consuming and too costly.

4.1. Closing the Gap

Bringing the language industry into the research fold and vice-versa is a win-win situation for both. The development of language technology in a multi-billion dollar language industry with an annual growth rate of almost 5%[7] is extremely lucrative for those whose business is language, and if the research community can demonstrate visible, profitable and concrete technological innovation and breakthroughs in application scenarios, they will make a good case for significantly more funded research in the field.

Quality evaluation development that incorporates the needs of both communities can provide the necessary impetus for more collaborative efforts and promote a greater level of understanding of the work each group does. Some open-source tools such as translate5[8] are now beginning to understand these parallels and are developing environments that combine the business features required by industry with the scientific features required by research. The goal is to turn translate5 into a flexible repository and data curation tool for MT research going beyond the functionality that can be provided by open resource exchange and sharing facilities such as META-SHARE (Burchardt et al., 2016).

4.2. Single Environment for Multiple Objectives

A holistic environment that combines quality evaluation requirements for professional translation and machine translation output in both business and research applications and offers flexible tool integration for different evaluation scenarios will provide the foundation for novel and groundbreaking research in improving machine translation quality. Incorporating the linguistic and language-related knowledge of industry experts into machine translation research can uncover previously unattainable information that is vital to the improvement process.

Until now, the language industry has relied primarily on human resources to manually fix issues in the machine translation output to achieve a suitable level of quality. This process does not address, help understand, or permanently remedy underlying errors. It is not that the errors are not understood or that the user does not want to apply the information to improve the next translation. The system, tools

[7] http://www.pangeanic.com/knowledge_center/size-of-the-translation-industry/

[8] http://www.translate5.net

and workflow do not support the incorporation of this type of information, so the information is not collected despite its valuable potential. The heterogenous translation environments and large number of quality standards complicate matters.

The research community, on the other hand, has focused much of its quality evaluation effort on improving the scores of automated metrics, sometimes based on reference translations completed by human resources, other times based on rankings and other forms of overall evaluations. Rarely does the feedback from linguistic experts find its way into ongoing research, and manual tasks such as annotation or error categorization are seen as too costly and ineffective. Without some of this information, it is difficult for the research community to see the benefits of applying it. Moreover, much of the research performed on its own is related to and can profit from research performed elsewhere.

A single, common environment that can connect all of these constituencies with each other, allow them to share information and results, experiment with data to which they would otherwise have no or little access can facilitate a level of communication that promotes cooperation and innovation. It can provide industry with a standardized platform that supports the import and export of files in any format, the definition of flexible quality metrics using MQM and other tools, the annotation and post-editing of machine translation for improvement cycles. It will make the efforts of the research community more accessible and comprehensible,

In turn, the research community will benefit from the work performed by industry users, making the quid pro quo collaboration on a unified platform affordable. It will have quick and easy access to data and results of other research users in an endless respository and the ability to plug-and-play almost any of the 2,725 language resources on META-SHARE.

5. Conclusions

The development of a holistic environment for translation quality evaluation that encompasses the requirements of both the research community and the language industry can have a significant positive impact on the future of language technology, in particular machine translation. It can provide the foundation for closer collaboration between the constituencies most interested in improving machine translation and secure the future of language technology and the translation industry.

Acknowledgements

This article has received support from the EC's Horizon 2020 research and innovation programme under grant agreements no. 645452 (QT21) and no. 645357 (CRACKER). We thank the anonymous reviewers for their valuable comments.

References

Autodesk. (2011). Translation and Post-Editing Productivity. In *http://translate.autodesk.com/productivity.html*.

Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.

Burchardt, A., Harris, K., Rehm, G., and Uszkoreit, H. (2016). Towards a systematic and human-informed paradigm for high-quality machine translation. In *Translation evaluation – From fragmented tools and data sets to an integrated ecosystem, LREC 2016 workshop*.

Doherty, S., Gaspari, F., Groves, D., van Genabith, J., Specia, L., Burchardt, A., Lommel, A., and Uszkoreit, H. (2013). Mapping the industry i: Findings on translation technologies and quality assessment. QTLaunchpad, FP7 funded by the European Union, Grant number 296347.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2016). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28, 1.

Koby, G. S., Fields, P., Hague, D., Lommel, A., and Melby, A. (2014). Defining translation quality. *Revista Tradumàtica*, Traducció i qualitat(Número 12), December.

Papineni, K., Roukos, S., Ward, T., and jing Zhu, W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, PA.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Zaret, A. (2016). A quality evaluation template for machine translation. *Translation Journal*, January.