Book Review

Quality Estimation for Machine Translation

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold
(University of Sheffield and Federal University of Technology, Paraná)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 39), 2017, xvi+121 pp; paperback, ISBN 978-1-68173-373-9, 64.95; ebook, ISBN 978-1-68173-374-6; doi:10.2200/S00584ED1V01Y201805HLT039

Reviewed by
François Yvon
LIMSI, CNRS, Université Paris-Saclay

Many natural language processing tasks aim to generate a human readable text (or human audible speech) in response to some input: Machine translation (MT) generates a target translation of a source input; document summarization generates a shortened version of its input document(s); text generation converts a formal representation into an utterance or a document, and so on. For such tasks, the automatic evaluation of the system's performance is often performed by comparison to a reference output, deemed representative of human-level performance.

Evaluation of MT is a typical illustration of this approach and relies on metrics such as BLEU (Papineni et al. 2002), Translation Edit Rate (TER) (Snover et al. 2006), or METEOR (Banerjee and Lavie 2005) implementing various string comparison routines between the system output and the corresponding reference(s). This strategy has the merit of making evaluation fully automatic and reproducible. Preparing human translation references is, however, a costly process, which requires highly trained experts; it is also prone to much variability and subjectivity. This implies that the failure to match the reference does not necessarily entail an error of the system. Reference-based evaluations are also considered too crude for many language pairs and tend to only evaluate the system's ability to reproduce one specific human annotation. Organizers of shared tasks in MT have therefore abandoned reference-based metrics to compare systems and resort to human judgments (Callison-Burch et al. 2008).

The book by Specia, Scarton, and Paetzold surveys an alternative approach to automatic evaluation of MT, *Quality Estimation* (QE). In essence, QE aims to move away from human references and to evaluate a generated text based only on automatically computed features. QE was initially proposed in the context of Automatic Speech Recognition (ASR) systems, an area where much of the foundational work has been performed (Jiang 2005). As explained in the introductory chapter, QE for MT also has many applications and has emerged in the last decade as a very active subfield of MT, with its own evaluation campaigns and metrics. In a nutshell, QE predicts the quality score or quality label of some target fragment, produced in response to some source text. Assuming that texts annotated with their quality level are available, QE is usually cast as a supervised machine learning task. QE for MT needs to simultaneously take two dimensions into account: (a) Is the proposed output appropriate for the input data?

doi:10.1162/COLI_r_00352

© 2019 Association for Computational Linguistics
Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license

and (b) Is the generated text grammatically correct? Dimension (a) is usually associated with the concept of *adequacy* with respect to the input signal, whereas dimension (b) is associated with the correctness or *fluency* of the output target fragment. Correctness can be defined at various levels of granularity, depending on the size of the output chunk: Smaller chunks need to contain the right words or to be syntactically correct; larger chunks, in addition, need to contain valid discourse relationships and to display lexical cohesiveness.

Specia and her colleagues address all these issues, and many more, in this book, where they survey QE for MT from an increasingly larger perspective: first words and phrases, then sentences, and finally complete documents. The last chapter widens the perspective, surveying QE for related NLP tasks, such as text simplification and generation.

After the short introduction, the second chapter covers QE for MT at the subsentential level, adopting a fixed outline that will also be used in the subsequent chapters. It successively presents the main applications, the targeted labels, the features, evaluation metrics, and some selected state-of-the-art approaches. This task nicely illustrates the difference between reference-based evaluation and QE: Evaluating the correctness of a hypothetical isolated word or phrase with respect to a reference translation would only be possible in rare cases of technical terms or named entities. Yet, it is possible to devise useful quality indicators at the word or phrase levels using, for instance, traces of the translation post-editor's work. Post-edition produces an annotation of which words are "correct" and kept in the revised version, and which are "wrong" and need to be discarded, replaced, or moved. Building QE systems using such labels amounts to training a sequence labeling system, for which many computational architectures, many features, and obvious evaluation metrics (e.g., label accuracy) are readily available. As discussed at length in the last section, more complex views of this task have proven useful to establish state-of-the-art results. For instance, Automatic Post-Edition–based QE predicts quality labels using an automatic post-edition of the output.

Chapter 3 is devoted to sentence-level QE, and follows the same outline as the previous chapter. Sentence-level QE is useful for many downstream applications, as evidenced by the large choice of labels and scores that can be targeted. It is still possible to use discrete labels such as "usable/useless sentence," or richer measures of the post-editing effort. Considering complete sentences makes it possible to also predict continuous values, such as the TER score, or the post-editing time. For each of these settings, off-the-shelf learning tools are available, as well as appropriate evaluation metrics. Note that such tasks are hard, probably as hard as MT itself: Indeed, any system successful for this task could effectively re-rank the output of an MT system and readily yield improved automatic translations. Working at the sentence level also encourages the use of very diverse sets of features, which the authors have taken the burden to list exhaustively: In addition to the obvious syntactic features for evaluating fluency, the authors also describe features that detect translation difficulties in the source (complexity indicators), extract information from the internals of the MT (confidence indicators), and globally evaluate the source and target alignment (translation features).

Chapter 4 covers document-level QE, which may provide users with a useful score, both for gisting applications and also in post-edition or translation revision scenarios. This is a new and difficult topic, as acknowledged by the small number of studies on these issues. One of the main challenges concerns the definition of metrics that could be used to define objective training functions: Possible sources of inspiration come from automatic text comprehension metrics or variants of post-edition effort. Document level QE also needs features that measure document-level properties of a text such

as discourse-structure appropriateness or lexical cohesion, an area where techniques borrowed from the statistical and neural text-mining literature such as topic models can be very useful. Much, however, remains to be done in this domain.

QE for other language generation applications is discussed in Chapter 5, which notably covers Text Simplification, Automatic Text Summarization, Grammatical Error Correction, and Natural Language Generation. With the exception of QE for ASR, which is already well documented, developing QE for these tasks is relatively new, as acknowledged by the small amount of prior work. Considering multiple applications poses new questions regarding quality measures and techniques that can be used to approximate them. For each task, the authors follow the same organization as in the previous chapters, which altogether puts considerable emphasis on the descriptions of features, as many features are useful in more than one task. This chapter is nonetheless very informative regarding the development of QE methods for systems generating a text output.

The concluding chapter discusses some directions for future research, in the light of the advent of a new generation of neural machine translation systems: On the one hand, neural machine translation outputs are quite different from statistical machine translation outputs, suggesting that QE methods need to evolve and take this difference into account; on the other hand, state-of–the-art QE systems are using neural components and need to be improved in their ability to train with scarce annotated data and to adapt to new domains and language pairs. This concluding section also plays the role of an appendix, as it includes a very complete list of existing resources and software packages for quality estimation of MT output. This supplementary material, and the bibliography that follows, will be of great help for readers willing to re-implement the systems described in this book or to develop new ideas.

In summary, this book discusses problems whose significance (at least for MT) is increasing with the general quality of machine translation output: With MT becoming more widespread and useful, it becomes necessary to inform users with indications of the cases where MT can be relied on—or not. Due to the invaluable expertise of the authors, who have been directly involved in the development of this field, the coverage of the recent literature on QE for MT (and on QE in general) is extremely thorough, which makes this book a must-read for any NLP practitioner willing to develop QE systems. The reading of this quite technical book, however, requires a solid working knowledge of machine translation and previous exposure to both automatic language analysis and to machine learning methodologies. For the sake of accessibility to a larger audience, adding definitions of basic concepts of the domain and a glossary should be considered in the second edition of the book.

References

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, pages 65–72, Ann Arbor, MI.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, OH.

Jiang, Hui. 2005. Confidence measures for speech recognition: A survey. *Speech Communication*, 45:455–470.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John

Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231, Boston, MA.

François Yvon is a senior CNRS researcher at LIMSI, Université Paris-Saclay, in Orsay, France. His research interests are mainly in natural language processing, including text mining, multi-lingual language processing, and machine translation. His e-mail address is `francois.yvon@limsi.fr`.