Quality Estimation for Language Output Applications

Carolina Scarton and **Gustavo Henrique Paetzold** and **Lucia Specia**
Department of Computer Science
University of Sheffield, UK
{c.scarton,g.h.paetzold,l.specia}@sheffield.ac.uk

Description

Quality Estimation (QE) is the task of predicting the quality of the output of Natural Language Processing (NLP) applications without relying on human references. This is a very appealing method for language output applications, i.e. applications that take as input a text and produce a different text as output, for example, Machine Translation, Text Summarisation and Text Simplification. For these applications, producing human references is time consuming and expensive. More important, QE enables quality assessments of the output of these applications *on the fly*, making it possible for users to decide whether or not they can rely and use the texts produced. This would not be possible with evaluation methods that require datasets with gold standard annotations. Finally, QE can predict scores that reflect how good an output is for a given purpose and, therefore, is considered a task-based evaluation method. The only requirement for QE are data points with quality scores to train supervised machine learning models, which can then be used to predict quality scores for any number of unseen data points. The main challenges for such a task rely on devising effective features and appropriate labels for quality at different granularity levels (words, sentences, documents, etc.). Sophisticated machine learning techniques, such as multi-task learning to model biases and preferences of annotators, can also contribute to making the models more reliable.

Figure 1 illustrates a standard framework for QE during its training stage. Features for training the QE model are extracted from both source (original) and target (output) texts (and optionally from the system that produced the output). A QE model can be trained to predict the quality at different granularity levels (such as words, sentences and documents) and also for different purposes. Therefore the input text, the features, the labels and the machine learning algorithm will depend on the specificities of the task variant. For example, if the task is to predict the quality of machine translated sentences for post-editing purposes, a common quality label could be post-editing time (i.e. the time required for a human to fix the machine translation output), while features could include indicators related to the complexity of the source sentence and the fluency of the target sentence.

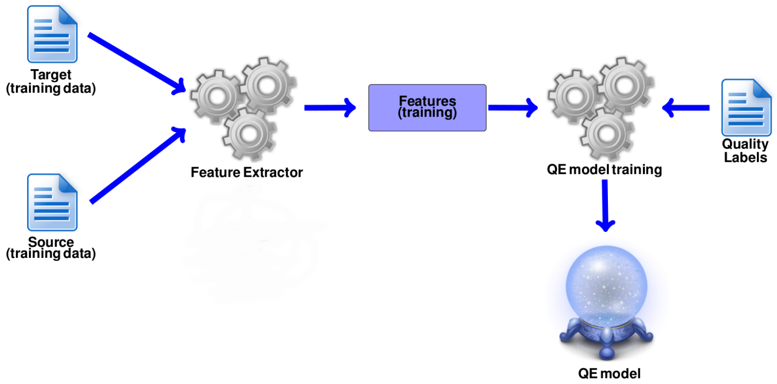

Figure 1: QE training framework.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Tutorial Abstracts,
pages 14–17, Osaka, Japan, December 11-17 2016.

Figure 2 illustrates the usage of a trained QE model. Unseen unlabelled data is the input for such a stage. The same features that were used to train the QE model are extracted from these instances and quality predictions are then produced by the model.

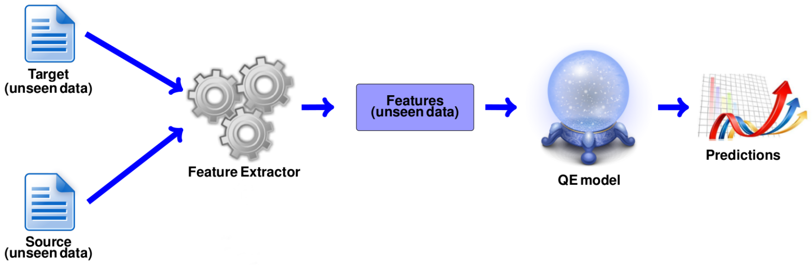

Figure 2: QE model prediction.

QE is a reasonably new field, but over the last decade has become particularly popular in the area of **Machine Translation (MT)**. With the goal of providing a prediction on the **quality of a machine translated text**, QE systems have the potential to make MT more useful in a number of scenarios, for example, improving post-editing efficiency by filtering out segments which would require more effort or time to correct than to translate from scratch (Specia, 2011), selecting high quality segments (Soricut and Echihabi, 2010), selecting a translation from either an MT system or a translation memory (He et al., 2010), selecting the best translation from multiple MT systems (Shah and Specia, 2014), and highlighting words or phrases that need revision (Bach et al., 2011).

Sentence-level QE represents the vast majority of existing work. It has been addressed as a supervised machine learning task using a variety of algorithms to train models from examples of sentence translations annotated with quality labels (e.g. 1 to 5 *likert* scores). This prediction level has been covered in shared tasks organised by the Workshop on Statistical Machine Translation (WMT) annually since 2012 (Callison-Burch et al., 2012; Bojar et al., 2013; Bojar et al., 2014; Bojar et al., 2015; Bojar et al., 2016). While standard algorithms can be used to build prediction models, key to this task is work of feature engineering.

Word-level QE has also been receiving significant attention. It is seemingly a more challenging task, where a quality label is to be produced for each target word. An additional challenge for this level is the acquisition of sizeable training sets. Significant efforts have been made (including four years of shared tasks at WMT), leading to an increase in interest in word-level QE over the years. An application that can benefit from word-level QE is spotting errors (incorrect words) in a post-editing/revision scenario. A recent variant of this task is quality prediction at the level of phrases (Logacheva and L.Specia, 2015; Blain et al., 2016), where a phrase can be defined in different ways, e.g. using the segmentation from a statistical MT decoder in WMT16 (Bojar et al., 2016).

Document-level QE has received much less attention than the other levels. This task consists in predicting a single quality label for an entire document, be it an absolute score (Scarton and Specia, 2014) or a relative ranking of translations by one or more MT systems (Soricut and Echihabi, 2010). It is most useful for *gisting* purposes, where post-editing is not an option. Two shared tasks on document-level QE were organised at WMT15 and WMT16. An open research question when it comes to document-level QE is to define effective quality labels for entire documents (Scarton et al., 2015).

A few QE frameworks have been proposed in the last couple of years, namely (Gonzàlez et al., 2012; Specia et al., 2013; Servan et al., 2015; Logacheva et al., 2016; Specia et al., 2015). QUEST++ (Specia et al., 2015)[1] is the most widely used one. It is a significantly refactored and expanded version of QUEST (Specia et al., 2013). It has been used as the official baseline system during all editions of the WMT shared task on QE (WMT12-WMT16) and is often the starting point upon which other participants build their systems, particularly for feature extraction. It has two main modules: feature extraction and

[1] https://github.com/ghpaetzold/questplusplus

machine learning. The feature extraction module can operate at sentence, word and document-level and includes the extraction of shallow and linguistic-motivated features. The machine learning module provides wrappers for various algorithms in SCIKIT-LEARN (Pedregosa et al., 2011), in addition to a few implementations of stand-alone algorithms such as Gaussian Processes for regression.

QE of MT will be the focus of this tutorial, but we will also introduce related work in applications such as Text Summarisation and discuss how the same framework could be applied or adapted to other language output applications.

Tutorial Structure

Theoretical aspects of QE (1h30) In the first part of this tutorial we will introduce the task of QE, show the standard framework for it and describe the three most common levels of prediction in QE for MT. We will also introduce various ways in which different kinds of Neural Networks can be employed in QE. Challenges and future work for each level will be discussed. We will cover related work for NLP applications other than MT, including ideas on how to adapt the current QE pipeline. In addition, examples of uses of QE in research and industry will be illustrated.

Hands-on QUEST++ (1h00) The second part of the tutorial will cover a hands-on QUEST++ activity, showing how to install and run it for examples available at all prediction levels. We will describe the two modules of this framework: feature extractor (implemented in Java) and machine learning (implemented in Python). We will guide participants to add an example of linguistic feature to QUEST++ using external resources, showing the interaction between classes and configuration files. We will also show how to write a wrapper for a new machine learning algorithm from SCIKIT-LEARN .

More information will be made available at the tutorial's website: `http://staffwww.dcs.shef.ac.uk/people/C.Scarton/qe-tutorial/`.

Instructors

Carolina Scarton is a PhD student and a Research Assistant at the University of Sheffield, UK, being supervised by Professor Lucia Specia. The topic of her thesis is on document-level assessment for QE of MT. More specifically, her research focuses on how to assess machine translated documents in order to build QE models at document level and the development of features for document-level QE (including the contribution on document-level QE for QuEst++). She has published several papers in international conferences in topics related to QE of MT and organised the document-level task on WMT15[2] and WMT16[3] QE shared tasks. Additionally, her research topics also include Readability Assessment, Text Simplification and Language Acquisition. Carolina received a Master degree in Computer Science from the University of São Paulo, Brazil, in 2013.
webpage: `http://www.dcs.shef.ac.uk/people/C.Scarton/`

Gustavo Henrique Paetzold is a Research Assistant at the University of Sheffield, UK, with a Ph.D. in Computational Linguistics. His main areas of expertise are Text Adaptation and Quality Estimation. Throughout the past few years, Gustavo has published several contributions to QE in international conferences, and is one of the main contributors to the QuEst++ framework, which will be feature in this tutorial. He has also experience in developing QE solutions for industry, given his brief collaboration with Iconic Translation Machines Ltd.
webpage: `https://gustavopaetzold.wordpress.com`

Lucia Specia is a Professor of Language Engineering and a member of the Natural Language Processing group at the University of Sheffield, UK. Her main areas of research are MT, Text Adaptation, and Quality Evaluation and Estimation of language output applications. Prof Specia is the recipient of an ERC Starting Grant on Multimodal Machine Translation (2016-2021) and is currently involved in various other funded research projects, including the European initiatives QT21 (Quality Translation 21) and

[2] `http://www.statmt.org/wmt15/quality-estimation-task.html`

[3] `http://www.statmt.org/wmt16/quality-estimation-task.html`

Cracker (Cracking the Language Barrier). She has published over 100 research papers in peer-reviewed journals and conference proceedings and organised a number of workshops in the area of NLP. She has given six tutorials on topics related to MT.
webpage: `www.dcs.shef.ac.uk/people/L.Specia/`

References

N. Bach, F. Huang, and Y. Al-Onaizan. 2011. Goodness: a method for measuring MT confidence. In *ACL11*.

F. Blain, V. Logacheva, and L. Specia. 2016. Phrase level segmentation and labelling of machine translation errors. In *LREC16*.

O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. 2013. Findings of the 2013 Workshop on SMT. In *WMT13*.

O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna. 2014. Findings of the 2014 Workshop on SMT. In *WMT14*.

O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, M. Huck, C. Hockamp, P. Koehn, V. Logacheva, C. Monz, M. Negri, M. Post, C. Scarton, L. Specia, and M. Turchi. 2015. Findings of the 2015 Workshop on SMT. In *WMT15*.

O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Neveol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *WMT16*.

C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. 2012. Findings of the 2012 Workshop on SMT. In *WMT12*.

M. Gonzàlez, J. Giménez, and L. Màrquez. 2012. A Graphical Interface for MT Evaluation and Error Analysis. In *ACL12*.

Y. He, Y. Ma, J. van Genabith, and A. Way. 2010. Bridging SMT and TM with translation recommendation. In *ACL10*.

V. Logacheva and L.Specia. 2015. Phrase-level quality estimation for machine translation. In *IWSLT15*.

V. Logacheva, C. Hokamp, and L. Specia. 2016. Marmot: A toolkit for translation quality estimation at the word level. In *LREC16*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12.

C. Scarton and L. Specia. 2014. Document-level translation quality estimation: exploring discourse and pseudo-references. In *EAMT14*.

C. Scarton, M. Zampieri, M. Vela, J. van Genabith, and L. Specia. 2015. Searching for Context: a Study on Document-Level Labels for Translation Quality Estimation. In *EAMT15*.

C. Servan, N.-T. Le, N. Q. Luong, B. Lecouteux, and L. Besacier. 2015. An Open Source Toolkit for Word-level Confidence Estimation in Machine Translation. In *IWSLT15*.

K. Shah and L. Specia. 2014. Quality estimation for translation selection. In *EAMT14*.

R. Soricut and A. Echihabi. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *ACL10*.

L. Specia, K. Shah, J. G. C. de Souza, and T. Cohn. 2013. Quest - a translation quality estimation framework. In *ACL13*.

L. Specia, G. H. Paetzold, and C. Scarton. 2015. Multi-level Translation Quality Prediction with QuEst++. In *ACL-IJCNLP15 - System Demonstrations*.

L. Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *EAMT11*.