Final Activity Report

Predicting Relevance and Quality of News Translation

https://github.com/sheffieldnlp/deepQuest

Julia Ive, Lucia Specia, Fred Blain

University of Sheffield

1. Introduction

The goal of this project was to devise a Quality Estimation (QE) approach framed specifically in the context of Machine Translation (MT) of news articles. News are abundant and often multiple articles on the same topic are written in a given language. Automatically translating such articles could better enable news to reach foreign audiences, giving readers in a foreign language the opportunity to access articles in subjects other than those of local interest, and to access information delivered from different points of view. The scenario of interest is thus that of translation for gisting purposes. However, this type of content introduces important challenges to state of the art machine translation, which often results in far from perfect quality translations, and thus automatic quality estimation becomes paramount. One such a challenge is that – different from the usual and rather artificial scenario according to which MT is often developed and evaluated, where sentences are translated in isolation – news articles need to be translated as an entire piece of text, ensuring lexical cohesion, accurate referencing, etc. Another major feature in news articles that is disregarded when translating and evaluating this type of text, is that certain parts of an article are more important than others. News articles often include background sections, as well as text on related topics. When translating and evaluating such translations, intuitively one would expect the core parts of the article to be translated accurately, whereas that is not as important a requirement for the remaining text. Existing QE approaches are not adequate for this gisting scenario, which is very different from standard sentence-level estimation for post-editing purposes. We propose a novel approach that takes into account the peculiarities of the type of text, the nature of the task and its purpose: prediction will be generated for an article as a whole, taking into account the relevance of the article itself (for cases where multiple articles cover the same topic) and of its component sentences and words with respect to the news. This approach involves novel methods to extract relevance information from both original and translated articles at different levels (word, sentence, full article), and novel ways of modelling the prediction task as a multi-level problem.

The key research objectives we aimed to address in this project were:

• To investigate relevance information at multiple levels (article, word, segment), for original and translated articles.

• To devise ways to integrate relevance and adequacy indicators in a document-level QE pipeline.

• To develop a new toolkit for quality estimation and release it as open source.

Initially we had proposed to build on the QuEst++ framework. However, meanwhile novel neural approaches were proposed for the problem (e.g. the POSTECH approach), showing much better performance for word and sentence-level prediction. Therefore, we took a step back and first re-implemented the best existing approach for these levels (for which code was not available), then extended it for document-level. We also created a more light-weight version of it which can be trained much more efficiently.

The project started April 2017 and was concluded in March 2018 with the code released and a paper submitted and accepted for the COLING 2018 conference. In this report we briefly describe the activities developed. The final results are shown in the COLING 2018 paper (to appear).

2. Activities

QE for news articles translations for gisting constitutes a very different scenario from the usual sentence-level prediction of post-editing effort: certain sentences of an article are more important than others, and it is therefore important that these sentences are translated accurately. To identify the importance of each sentences within a given document, we usually rely on Term Frequency - Inverse Document Frequency (tf-idf), a technique used in text summarization to extract a few sentences that best summarized the document. Tf-idf is computed at word-level, and the relevance score of each sentence corresponds to the sum of each tf-idf scores of the words in that sentence. The following main activities were developed in the project.

2.1 Open-source version of SOTA QE model

Based on the results of the 2017 edition of the QE Shared task at WMT (Bojar et al., 2017), the POSTECH architecture, developed by (Kim et al., 2017), is considered as the state-of-the-art for both word and sentence-level predictions. However, the source code of such model was not available, and so our first step was to implement an open-source version of this architecture. Our objective was two-fold: i) extend its architecture to document-level; ii) allow our work to be reproducible and used by the research community. Another reason of choosing to implement this particular approach is the fact that POSTECH is a stacked

architecture (i.e. a representation at a certain level is built from representations at a lower level), which makes it relevant for the addition of a document-level movel in this project.

The POSTECH architecture consists of an encoder-decoder Recurrent Neural Network (RNN) (so called *predictor*), which predicts words along with context representation from an attention mechanism (Bahdanau et al., 2015), which are then used as input for an unidirectional RNN (so called *estimator*) that produces quality estimates for words, phrases and sentences. We based our implementation on the NMT-Keras library, an open-sourced library for neural machine translation, written in Python. We released our code in the form of an open-sourced framework for neural-based QE, called “deepQuest”. This framework is freely available online[1], and one can use it to produce quality estimates based on the POSTECH model.

2.2 Light-weight stacked architecture

Despite its strong performance at predicting quality estimates at different levels (word, phrase and sentence), the POSTECH architecture is very complex, time and resource consuming to train. As an alternative, we implemented a new, lighter, stacked architecture that uses bi-directional Recurrent Neural Networks, so-called *BiRNN*. In this approach, we train independently two bi-directional RNNs (one for the source sentence, and one for the machine translation). For word-level QE, the two representations are concatenated afterwards. For sentence-level QE on the other hand, we do not represent sentences as simple aggregations of word-level representations: this representation should reflect some importance of the words in that sentence. Thus, we need a certain weighting for each of those word-level representations, which is provided through an attention mechanism.

The source code of our BiRNN approach has also been released and documented in our open source framework deepQuest.

2.3 Document-level predictions

Similarly to POSTECH, our BiRNN architecture is suitable for both word and sentence-level predictions, but did not allow document-level predictions. Both of those models being stacked architecture, we expanded both models with a document-level representation built from representations at sentence-level.

This is done with a bi-directional RNN, used as an encoder. RNNs have been successfully used for document representation (Lin et al., 2015) and consequently applied to a series of downstream tasks such as topic labelling, summarization, and question answering (Li et al., 2015; Yang et al., 2016).

1 https://github.com/sheffieldnlp/deepQuest

Similarly to the sentence-level representations with our BiRNN model, our assumption is that we should not represent a document as the simple aggregation of sentence-level QE representations, but instead should reflect the importance of each sentence within that document. Therefore, we also use an attention mechanism to learn the weights of each sentence.

2.4 Experiments

The first step of our experiments was to determine the quality label to predict for the automatic translation of a news article, which is inspired by the notion of relevance we defined above.

While in machine translation sentences are usually translated and evaluated in isolation, we attempt to predict the quality of a news article as an entire piece of text, to evaluate lexical cohesion, accurate referencing, etc. To obtain a document-level score that considers the relevance of each of its sentences, we rely on a variant of the popular BLEU score. We follow Chen (2014)'s recommendation that a weighted average of sentence-level BLEU (wBLEU) scores (where the weight is the length of the sentence) achieved a better correlation with human judgement than the original IBM corpus-level BLEU. We also follow conclusions by Turchi, Steinberger and Specia (2012), and change the weight of each sentence in wBLEU by the relevance score of that sentence within the document (i.e. the sum of the tf-Idf scores of the words in that sentence). We call this variant tBLEU.

The second step was to find enough data to train our models, as we first considered using the document-level QE dataset by (Graham et al., 2017), however, the small number of documents (62) and language pairs (English-to-Spanish only) made this resource less applicable for this work.

Instead, we opted for all submissions at the WMT News shared tasks for various years, a task where each participating system is required to translate a set of News documents. This gives us access to a large set of language pairs, as well as provides us with a range of different translation quality. We collected system submissions from WMT 2008 up to 2017 for four language pairs: German-English (DE–EN, 14,640 documents for 2008-2017, excluding 2010), and English–Spanish (EN–ES, 6,733 documents for 2008-2013, excluding 2010), English–French (EN–FR, 11,537 documents for 2008-2014) and English-Russian (EN–RU, 6,996 documents for 2013-2017). For each language pair, we consider either the full set of system submissions, or a filtered version of it composed only of both the best and the worst performing systems for each year. The filtering was done based on the overall BLEU score achieved by each system as reported on matrix.statmt.org. Our intuition is that by considering only the extreme quality cases we would make our data, while smaller, easier to discriminate.

We also report on the official metrics of the WMT QE shared task (MAE and Pearson correlation). Our results shown that our BiRRN model particularly correlate with the tBLEU scores. We tend to attribute this to the fact that our architecture builds document-level representation from sentence-level representations, which in turn depend on word representations. The tBLEU reflects this hierarchy in the most consistent way as those document-level scores depend directly on semantic importance of words they contain.

Details on the experiments are reported in Ive, Blain and Specia (2018).

To facilitate the reproduction of our work, as well as encourage the development of new approaches for document-level QE, we make the all data processing pipeline available, along with the source code of our QE models: https://github.com/sheffieldnlp/deepQuest.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

Chen, Boxing, and Colin Cherry. "A systematic comparison of smoothing techniques for sentence-level BLEU." *Proceedings of the Ninth Workshop on Statistical Machine Translation*. 2014.

Graham, Yvette, et al. "Improving Evaluation of Document-level Machine Translation Quality Estimation." *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Vol. 2. 2017.

Bojar, Ondřej, et al. "Findings of the 2017 conference on machine translation (wmt17)." *Proceedings of the Second Conference on Machine Translation*. 2017.

Ive, Julia, Blain, Fred, Specia, Lucia. “deepQuest: a Framework for neural-based Quality Estimation”. Proceedings of COLING. 2018 (to appear).

Kim, Hyun, Jong-Hyeok Lee, and Seung-Hoon Na. "Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation." *Proceedings of the Second Conference on Machine Translation*. 2017.

Lin, Rui, et al. "Hierarchical recurrent neural network for document modeling." *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015.

Turchi, Marco, Josef Steinberger, and Lucia Specia. "Relevance ranking for translated texts." *Proceedings of the 16th Annual Conference of the European Association for Machine Translation, number May*. 2012.

Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016.