Part III

Applications - Translation Quality Estimation

Chapter 8

Introduction to Translation Quality Estimation

The goal of quality estimation (QE) is to evaluate the quality of a translation without having access to a reference translation [180]. High-accuracy QE that can be easily deployed for a number of language pairs is the missing piece in many commercial translation workflows as they have numerous potential uses. They can be employed to select the best translation when several translation engines are available or can inform the end user about the reliability of automatically translated content. In addition, QE systems can be used to decide whether a translation can be published as it is in a given context, or whether it requires human post-editing before publishing or even translation from scratch by a human [181].

Quality estimation task is different from automatic machine translation evaluation [182]. Automatic machine translation evaluation approaches like BLEU [176] need the reference translation too in order to perform the machine translation evaluation. As aforementioned, quality estimation on the other hand, does not need the reference translation. Therefore, automatic MT evaluation approaches like BLEU [176], METEOR [178], LEPOR [183] and others can not be directly applied in the QE task. As a result, the solutions proposed for QE are

completely different from that of automatic machine translation evaluation.

The estimation of translation quality can be done at different levels: word/phrase-level sentence-level and sentence-level [184]. Word-level QE aims to spot words that need to be reviewed during the post editing process. It indicates which words from the source have been incorrectly translated in the target, and whether the words inserted between these words are correct. Sentence-level QE models provide a single score for each pair of source and target sentences. Sentence-level QE scores help to rank translations that are worth post-editing. Document-level QE, on the other hand, scores or ranks documents according to their quality for fully automated MT usage scenarios [184]. In recent years, word-level QE and sentence-level QE have been more popular among the community [180].

During the past decade there has been tremendous progress in the field of quality estimation, largely as a result of the QE shared tasks organised annually by the Workshops on Statistical Machine Translation (WMT), more recently called as the Conferences on Machine Translation, since 2012 [180, 185, 186, 187, 188, 189, 190, 191, 192]. First, they have provided annotated datasets which can be used to train QE models and to evaluate them. Second, The annotated datasets these shared tasks released each year have led to the development of many open-source QE systems.

As we discuss in Section 8.1, at present neural-based QE methods constitute the state of the art in quality estimation. However, these approaches are based on complex neural networks and require resource-intensive training. This resource-intensive nature of these deep learning based frameworks makes it expensive

to train QE models. Furthermore, these architectures require a large number of annotated instances for training, making the quality estimation task very difficult for low-resource language pairs. This nature of the current state-of-the-art QE systems has hindered their popularity in real world applications.

In this Part of the thesis, we address these problems by employing simple STS architectures we experimented in Part I of the thesis in quality estimation. We redefine the QE task as a crosslingual STS task and show that state-of-the-art STS architectures can be applied in QE task by just changing the input embeddings. As far as we know, this is the first study done on applying neural STS models to QE task.

In Chapter 9, we explore sentence-level QE with STS architectures. We evaluate their performance in recent sentence-level QE datasets comparing them with open-source QE tools like OpenKiwi and QuEst++. Finally we propose a new state-of-the-art QE method for sentence-level QE.

In Chapter 10, we expand this idea to word-level QE. We modify the output of the STS architecture minimally to predict word-level translation qualities. We evaluated their performance in recent word-level QE datasets comparing the results with open-source word-level tools like OpenKiwi and Marmot and we show that in majority of the datasets our simple word-level architecture outperforms other QE tools.

In Chapter 11, for the first time, we explore multilingual QE with state-of-the-art word-level and sentence-level QE methods. Furthermore, we evaluate multilingual QE in different training environments like zero-shot and few-shot.

Our findings in Chapter 11, would be beneficial for low resource languages.

The main contributions of this part of the thesis are as follows.

1. We propose two STS architectures based on Transformers to perform sentence-level QE. These architectures are simpler than the architectures available in OpenKiwi [181] and DeepQuest [184]. We evaluate them on both aspects of sentence-level QE on 15 language pairs and we show that the two architectures outperform the current state-of-the-art sentence-level QE frameworks like OpenKiwi [181] and DeepQuest [184].

2. We introduce a simple architecture to perform word-level quality estimation that predicts the quality of the words in the source sentence, target sentence and the gaps in the target sentence. We evaluate it on eight different language pairs which the word-level QE data was available and we show that the proposed architecture outperforms the current state-of-the-art word-level QE frameworks like Marmot [193] and OpenKiwi [181].

3. We propose multilingual learning for QE with the proposed architectures for sentence-level and word-level which is beneficial for low resource language pairs.

4. We provide important resources to the community. The code of the each chapter is bundled as an open-source QE framework and the pre-trained sentence-level and word-level QE models will be freely available to the community. The link to the relevent code and the models will be unveiled in the introduction section of the each chapter.

The remainder of this chapter is structured as follows. Section 8.1 explores the various methods that have been employed in sentence-level and word-level QE including previous research done with regard to STS in QE task. Section 8.2 discussed the various datasets we used in this part of the thesis. In Section 8.3 we show the main evaluation metrics we used for the sentence-level and word-level QE experiments in the following Chapters in Part III of the thesis. Chapter concludes with the conclusions.

8.1 Related Work

Before the neural network era, most of the quality estimation systems like QuEst [194] and QuEst++ [195] were heavily dependent on linguistic processing and feature engineering to train traditional machine-learning algorithms like support vector regression and randomised decision trees [194]. These features can be either extracted from machine translation system (*glass-box* features) or obtained from the source and translated sentences, as well as external resources, such as monolingual or parallel corpora (*black-box* features) [196]. For example QuEst [194] has 17 manually crafted features fed in to support vector regression algorithm. These 17 features consists of glass-box features such as *ratio of number of tokens in source and target segments*, *ratio of percentage of nouns/verbs in the source and target* etc. as well as black-box features such as *global score and relevant features of the SMT system*, *proportion of pruned search graph nodes* etc. too. In QuEst++ the number of features varies from 80 to 123 depending on the language pair [195]. QuEst [194], QuEst++ [195] and Marmot [193] can be

considered as the most popular traditional QE tools. QuEst [194] only supports sentence-level QE, Marmot [193] only supports word-level QE while QuEst++ [195] can support both word-level and sentence-level QE. Even though, they provided good results in early days, these traditional approaches are no longer the state-of-the-art. In recent years, neural-based QE systems have consistently topped the leader boards in WMT quality estimation shared tasks [181].

With the increasing popularity of word embeddings [15], neural networks based on word embeddings got popular in QE field too. They outperformed traditional QE systems and provided state-of-the-art results. For example, the best-performing system at the WMT 2017 shared task on QE was POSTECH, which is purely neural and does not rely on feature engineering at all [197]. POSTECH revolves around an encoder-decoder Recurrent Neural Network (RNN) (referred to as the 'predictor'), stacked with a bidirectional RNN (the 'estimator') that produces quality estimates. In the predictor, an encoder-decoder RNN model predicts words based on their context representations and in the estimator step there is a bidirectional RNN model to produce quality estimates for words, phrases and sentences based on representations from the predictor. To be effective, POSTECH requires extensive predictor pre-training, which means it depends on large parallel data and is computationally intensive [184]. The POSTECH architecture was later re-implemented in DeepQuest [184]. DeepQuest supports both word-level and sentence-level QE [184].

OpenKiwi [181] is another open-source QE framework developed by Unbabel. It implements four different neural network architectures QUETCH [198],

NuQE [199], Predictor-Estimator [197] and a stacked model of those architectures. Both the QUETCH and NuQE architectures have simple neural network models that do not rely on additional parallel data, but do not perform that well. The Predictor-Estimator model is similar to the POSTECH architecture and relies on additional parallel data. In OpenKiwi, the best performance for sentence-level quality estimation was given by the stacked model that used the Predictor-Estimator model, meaning that the best model requires extensive predictor pre-training and relies on large parallel data and computational resources.

The complex and resource intensive nature of these models create a few limitations when they are deployed in real-word scenarios. Therefore, in this study we propose to use simple STS architectures in QE. Over the years there have been a few attempts to integrate semantic similarity into QE. Specia et al. [200] bring semantic information into QE task in order to address the problem of meaning preservation in translation. The authors integrate semantic similarity features to the QE model and improve the results of the QE task [200]. Biçici and Way [201] introduce the use of referential translation machines (RTM) for QE. RTM is a computational model for judging monolingual and bilingual similarity that achieves state-of-the-art results. This approach provided the best result in both sentence level and word-level tasks of WMT 2013 [186]. Furthermore in to this, Kaljahi et al. [202] and Camargo de Souza et al. [203] used syntactic and semantic information in quality estimation and are able to improve over the baseline when combining these features with the features of the baseline. Finally, in a different approach, Bechara et al. [204] use semantically similar sentences and their qual-

ity scores as features to estimate the quality of machine translated sentences. They show that this method can improve the prediction of machine translation quality for semantically similar sentences [204].

Even though, there are several studies done to integrate semantic similarity into QE, as far as we know, this would be the first study to employ state-of-the-art neural STS models in the QE task.

8.2 Datasets

All the datasets that we used in this part of the thesis are publicly available and were released in WMT quality estimation tasks in recent years [180, 191, 192]. This was done to ensure replicability of our experiments and to allow us to compare our results with the state of the art. Following Sections would describe the sentence-level and word-level QE datasets we experimented, separately.

8.2.1 Sentence-level QE

Sentence-level QE datasets that we used in this Part of the thesis can be categorised in to two main areas depending on the aspect they have been annotated; Human-mediated Translation Edit Rate (HTER) and Direct Assessment (DA). Most of the early datasets have been annotated on HTER. Very recently DA aspect too is getting popular in the QE community. We describe each of them in details, in the following section.

Predicting HTER The performance of QE systems has typically been assessed using the semiautomatic HTER. HTER is an edit-distance-based measure which

Language Pair	Source	MT system	Competition	train, dev, test size
De-En	Pharmaceutical	Phrase-based SMT	WMT 2018 [180]	25,963, 1,000, 1,000
En-Zh	Wiki	fairseq based NMT	WMT 2020 [192]	7,000, 1,000, 1,000
En-Cs	IT	Phrase-based SMT	WMT 2018 [180]	40,254, 1,000, 1,000
En-De	Wiki	fairseq based NMT	WMT 2020 [192]	7,000, 1,000, 1,000
En-De	IT	Phrase-based SMT	WMT 2018 [180]	26,273, 1,000, 1,000
En-Ru	Tech	Online NMT	WMT 2019 [191]	15,089, 1,000, 1,000
En-Lv	Pharmaceutical	Attention-based NMT	WMT 2018 [180]	12,936, 1,000, 1,000
En-Lv	Pharmaceutical	Phrase-based SMT	WMT 2018 [180]	11,251, 1,000, 1,000

Table 8.1: Information about language pairs used to predict HTER. The **Language Pair** column shows the language pairs we used in ISO 639-1 codes[1]. **Source** expresses the domain of the sentence and **MT system** is the Machine Translation system used to translate the sentences. In that column NMT indicates Neural Machine Translation and SMT indicates Statistical Machine Translation. **Competition** shows the quality estimation competition in which the data was released and the last column indicates the number of instances the train, development and test dataset had in each language pair respectively.

captures the distance between the automatic translation and a reference translation in terms of the number of modifications required to transform one into another. In light of this, a QE system should be able to predict the percentage of edits required in the translation. We used several language pairs for which HTER information was available: English-Chinese (En-Zh), English-Czech (En-Cs), English-German (En-De), English-Russian (En-Ru), English-Latvian (En-Lv) and German-English (De-En). The texts are from a variety of domains and the translations were produced using both neural and statistical machine translation systems. More details about these datasets can be found in Table 8.1 and in [180, 191, 192].

Predicting DA Even though HTER has been typically used to assess quality in machine translations, the reliability of this metric for assessing the performance

[1]Language codes are available in ISO 639-1 Registration Authority Website Online - https://www.loc.gov/standards/iso639-2/php/code_list.php

of quality estimation systems has been questioned by researchers [205]. The current practice in MT evaluation is the so-called Direct Assessment (DA) of MT quality [206], where raters evaluate the machine translation on a continuous 1-100 scale. This method has been shown to improve the reproducibility of manual evaluation and to provide a more reliable gold standard for automatic evaluation metrics [207].

We used a recently created dataset to predict DA in machine translations which was released for the WMT 2020 quality estimation shared task 1 [192]. The dataset is composed of data extracted from Wikipedia for six language pairs, consisting of high-resource English-German (En-De) and English-Chinese (En-Zh), medium-resource Romanian-English (Ro-En) and Estonian-English (Et-En), and low-resource Sinhala-English (Si-En) and Nepalese-English (Ne-En), as well as a Russian-English (En-Ru) dataset which combines articles from Wikipedia and Reddit [208]. These datasets have been collected by translating sentences sampled from source-language articles using state-of-the-art NMT models built using the fairseq toolkit [209] and annotated with DA scores by professional translators. Each translation was rated with a score from 0-100 according to the perceived translation quality by at least three translators [192]. The DA scores were standardised using the z-score. The quality estimation systems evaluated on these datasets have to predict the mean DA z-scores of test sentence pairs. Each language pair has 7,000 sentence pairs in the training set, 1,000 sentence pairs in the development set and another 1,000 sentence pairs in the testing set.

8.2.2 Word-level QE

Word-level QE annotations are not straight-forward like sentence-level QE annotations. They have been annotated for words in the target ('OK' for correct words, 'BAD' for incorrect words), gaps in the target ('OK' for genuine gaps, 'BAD' for gaps indicating missing words) and source words ('BAD' for words that lead to errors in the target, 'OK' for other words) [180]. To make it clearer, consider following source and target sentences from a English to German translation. Their word-level quality estimation labels would be followed by. Please note that *yellow* represents 'OK' and *orange* represents 'BAD' labels.

Source - *for example , you could create a document containing a car that moves across the Stage .*

Target - *Sie können beispielsweise ein Dokument erstellen , das ein Auto über die Bühne enthält .*

Source - *for example , you could create a document containing a car that moves across the Stage .*

Target - *<GAP> Sie <GAP> können <GAP> beispielsweise <GAP> ein <GAP> Dokument <GAP> erstellen <GAP> , <GAP> das <GAP> ein <GAP> Auto <GAP> über <GAP> die <GAP> Bühne <GAP> enthält <GAP> . <GAP>*

As you can see, all the words in source, all the words in target and all the *gaps* in the target have been annotated as 'OK' or 'BAD'. Most of the language pairs that are annotated for sentence-level HTER scores have word-level quality scores too. Therefore, for the word-level QE experiments, we used the same

language pairs we used for sentence-level HTER which is shown in Table 8.1. More information above word-level annotations are available on [180, 191, 192].

8.3 Evaluation

For evaluation, we used the same approach proposed in the WMT shared tasks, so that we can compare our results with the respective baselines and the best systems submitted in each shared task. Obviously sentence-level and word-level QE follow two different evaluation criteria which we explain in the following sections.

Sentence-Level QE Evaluation : Similar to STS evaluation in Part I of the thesis, Pearson's Correlation Coefficient (ρ) is the most popular evaluation metric in recent WMT sentence-level QE shared tasks [180, 191, 192]. Since in both HTER and DA, the gold labels are continues and the models need to predict a continues value, it makes sense to employ Pearson's Correlation Coefficient as the evaluation metric. A QE model with a Pearson's Correlation Coefficient close to 1 indicates that the predictions of that model and gold labels have a strong positive linear correlation and therefore, it is a good model to predict sentence-level QE.

Word-level QE Evaluation : The primary evaluation metric for word-level QE is the multiplication of F1-scores for the OK and BAD classes, denoted as $F1_{MULTI}$ [180, 191, 192]. Standard equation for F1 score is shown in Equation 8.1 where TP, TN, FP and FN are True Positive, True Negative, False Positive and

False Negative respectively. This F1 score is calculated for OK and BAD classes seperately and then, they are multiplied to get the $F1_{MULTI}$ as shown in Equation 8.2.

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{8.1}$$

$$F1_{MULTI} = F1_{OK} \times F1_{BAD} \tag{8.2}$$

Prior to WMT 2019, $F1_{MULTI}$ score was calculated separately for words in source ($F1_{MULTI}$ $Source$), words in the target ($F1_{MULTI}$ $Target$) and gaps in the target ($F1_{MULTI}$ $GAPS$) [180], while after WMT 2019 [191, 192] they produce a single result for target gaps and words named as "$F1_{MULTI}$ $Target$" alongside with "$F1_{MULTI}$ $Source$". We follow the same approach. A QE model with a $F1_{MULTI}$ score close to 1 indicates that the predictions of that model and gold labels are similar and therefore, it is a good model to predict word-level QE.

8.4 Conclusion

Quality estimation is an important component in making machine translation useful in real-world applications, as it is aimed to inform the user on the quality of the MT output at test time. This process can be done on different levels such as word-level, sentence-level and document-level. QE shared tasks organised annually by WMT has increased the popularity of QE among MT community by leading the development of standard datasets covering variety of languages and

domains. These QE shared tasks have further contributed to the development of evaluation measures in QE. We followed the same evaluation measures; Pearson's Correlation Coefficient for sentence-level and $F1_{MULTI}$ for word-level.

The annotated datasets these shared tasks released each year have led to the development of many open-source QE systems. Similar to other NLP fields, most of the early QE approaches like QuEst [194], QuEst++ [195] and Marmot [193] were also based on traditional machine learning and involved heavy feature engineering. However, they no longer provide competitive results. Current state-of-the-art in QE is neural models. Following this, many open-source neural QE frameworks such as OpenKiwi [181] and DeepQuest [184] have been created. However, these neural QE architectures are complex and need a lot of computing resources to train a QE model which we have identified as a major limitation in them. To address this weakness, we propose to redefine the QE task as a crosslingual STS task and apply STS architectures we experimented in Part I of the thesis to QE which are considerably simpler than current state-of-the-art QE models.

In the next few chapters we will be exploring STS architectures in QE task. First in Chapter 9, the STS architectures will be applied in sentence-level QE and then in Chapter 10, these architectures will be extended to word-level QE. Finally in Chapter 11, we explore multilingual QE with the proposed STS architectures.