USFD's Phrase-level Quality Estimation Systems

Varvara Logacheva, Frédéric Blain and **Lucia Specia**
Department of Computer Science
University of Sheffield, UK
{v.logacheva, f.blain, l.specia}@sheffield.ac.uk

Abstract

We describe the submissions of the University of Sheffield (USFD) for the phrase-level Quality Estimation (QE) shared task of WMT16. We test two different approaches for phrase-level QE: (i) we enrich the provided set of baseline features with information about the context of the phrases, and (ii) we exploit predictions at other granularity levels (word and sentence). These approaches perform closely in terms of multiplication of F_1-scores (primary evaluation metric), but are considerably different in terms of the F_1-scores for individual classes.

1 Introduction

Quality Estimation (QE) of Machine Translation (MT) is the task of determining the quality of an automatically translated text without comparing it to a reference translation. This task has received more attention recently because of the widespread use of MT systems and the need to evaluate their performance on the fly. The problem has been modelled to estimate the quality of translations at the word, sentence and document levels (Bojar et al., 2015). Word-level QE can be particularly useful for post-editing of machine-translated texts: if we know the erroneous words in a sentence, we can highlight them to attract post-editor's attention, which should improve both productivity and final translation quality. However, the choice of words in an automatically translated sentence is motivated by the context, so MT errors are also context-dependent. Moreover, as it has been shown in (Blain et al., 2011), errors in multiple adjacent words can be caused by a single incorrect decision — e.g. an incorrect lexical choice can result in errors in all its syntactic dependants. The task of estimating quality at the phrase level aims to address these limitations of word-level models for improved prediction performance.

The first effort to estimate the quality of translated n-grams (instead of individual words) was described in (Gandrabur and Foster, 2003), but there the multi-word nature of predictions was motivated by the architecture of the MT system used in the experiment: an interactive MT system which did not translate entire sentences, but rather predicted the next n word translations in a sentence. An approach was designed to estimate the confidence of the MT system about the prediction and was aimed at improving translation prediction quality.

The phrase-level QE in its current formulation – estimation of the quality of phrases in a pre-translated sentence using external features of these phrases – was first addressed in the work of Logacheva and Specia (2015), where the authors segmented automatically translated sentences into phrases, labelled these phrases based on word-level labels and trained several phrase-level QE models using different feature sets and machine learning algorithms. The baseline phrase-level QE system used in this shared task was based on the results in (Logacheva and Specia, 2015).

This year's Conference on Statistical Machine Translation (WMT16) includes a shared task on phrase-level QE (QE Task 2p) for the first time. This task uses the same training and test data as the one used for the word-level QE task (QE Task 2): the set of English sentences, their automatic translations into German and their manual post-editions performed by professional translators. The data belongs to the IT domain. The training set contains 12,000 sentences, development and test sets — 1,000 and 2,000 sentences, respectively. For model training and evaluation, the words are la-

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 800–805,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

belled as “BAD” or “OK” based on labelling generated with the TERcom tool[1]: if an edit operation (substitution or insertion) was applied to a word, it is labelled as “BAD”; contrarily, if the word was left unchanged, it is considered “OK”. For the phrase-level task, the data was segmented also into phrases. The segmentation was given by the decoder that produced the automatic translations. The segments are labelled at the phrase level using the word-level labels: a phrase is labelled as “OK” if it contains only words labelled as “OK”; if one or more words in a phrase are “BAD”’, the phrase is “BAD” itself. The predictions are done at the phrase level, but evaluated at the word level: for the evaluation phrase-level labels are unrolled back to their word-level versions (i.e. if a three-word phrase is labelled as “BAD”, it is equivalent to three “BAD” word-level labels).

The baseline phrase-level features provided by the organisers of the task are *black-box* features that were originally used for sentence-level quality estimation and extracted using the QuEst toolkit[2] (Specia et al., 2015). While this feature set considers many aspects of sentence quality (mostly the ones that do not depend on internal MT system information and do not require language-specific resources), it has an important limitation when applied to phrases. Namely, it does not take into account the context of the phrase, i.e. words and phrases in the sentence, either before or after the phrase of interest. In order to advance upon the baseline results, we enhanced the baseline feature set with contextual information for phrases.

Another approach we experimented with is the use of predictions made by QE models at other levels of granularity: word level and sentence level. The motivation here is twofold. On the one hand, we use a wider range of features which are unavailable at the phrase level. On the other hand, the use of word-level and sentence-level predictions can help mitigate the uncertainty of phrase-level scores: there, a phrase is labelled as “BAD” if it has any number of “BAD” words, so “BAD” phrases can be of very different quality. We believe that information on the quality of individual words and the overall quality of a sentence can be complementary for phrase-level quality prediction.

The rest of the paper is organised as follows. We describe our context-based QE strategy in Section 2. In Section 3 we explain our approach to build phrase-level QE models using predictions of other levels. Section 4 reports the final results, while Section 5 outlines directions for future work.

2 Context-based model

The feature set used for the baseline system in the shared task considers various aspects of a phrase. It has features that allow to evaluate the likelihood of its source and target parts individually (e.g. probabilities of its source and target phrases as given by monolingual language models), and also the correspondences between the parts (e.g. the ratio of numbers of punctuation marks and words of particular parts of speech in the source and target sides of the phrase). However, this feature set does not take into account the words surrounding an individual phrase. This is explained by the fact that the feature set was originally designed for QE systems which evaluate the quality of automatic translations at the sentence level. Sentences in an automatically translated text are generally produced independently from each other, given that most MT systems cannot take extra-sentential context into account. Therefore, context features are rarely used for sentence-level QE.

2.1 Features

In order to improve the representation of phrases, we use a number of additional features (CONTEXT) that depend on phrases to the left and right of the phrase of interest, as well as the phrase itself. The intuition behind these features is that they evaluate how well a phrase fits its context. Here we list the new features and the values they can take:

- **out-of-vocabulary words (binary)** — we check if the source phrase has words which do not occur in a source corpus. The feature has value **1** if at least one of source words is out-of-vocabulary and **0** otherwise;

- **source/target left context (string)** — last word of the previous source/target phrase;

- **source/target right context (string)** — first word of the next source/target phrase;

- **highest order of n-gram that includes the first target word (0 to 5)** — we take the n-gram at the border between the current and

[1] http://www.cs.umd.edu/~snover/tercom/

[2] http://www.quest.dcs.shef.ac.uk/quest_files/features_blackbox

previous phrase and generate the combination of the first target word in the phrase and 1 to 4 words that precede it in the sentence. Let us denote the first word from the phrase $\mathbf{w}_{first}$ and the 4-grams from the previous phrase $p_{-4}p_{-3}p_{-2}p_{-1}$. If the entire 5-gram $p_{-4}p_{-3}p_{-2}p_{-1}\mathbf{w}_{first}$ exists in the target LM, the feature value is 5. If it is not in the LM, n-grams of lower order (from $p_{-3}p_{-2}p_{-1}\mathbf{w}_{first}$ to unigram w_{first}) are checked, and the feature value is the order of the longest n-gram found in the LM;

- **highest order of n-gram that includes the last target word (0 to 5)** — feature that considers the n-gram $\mathbf{w}_{last}p_1p_2p_3p_4$ (where $\mathbf{w}_{last}$ is the last target word of the current phrase and $p_1p_2p_3p_4$ is the opening 4-gram of the next feature) analogously to the previous feature;
- **backoff behaviour of first/last n-gram (0 to 1)** — backoff behaviour of n-grams $p_{-2}p_{-1}\mathbf{w}_{first}$ and $\mathbf{w}_{last}p_1p_2$, computed as described in (Raybaud et al., 2011).
- **named entities in the source/target (binary)** — we check if the source and target phrases have tokens which start with capital letters;
- **part of speech of the source/target left/right context (string)** — we check parts of speech of words that precede or follow the phrase in the sentence.

Some of these features (e.g. highest n-gram order, backoff behaviour, contexts) are used because they have been shown useful for word-level QE (Luong et al., 2013), others are included because we believe they can be relevant for understanding the quality of phrases.

We compare the performance of the baseline feature set with the feature set extended with context information. The QE models are trained using CRFSuite toolkit (Okazaki, 2007). We chose to train a Conditional Random Fields (CRF) model because it has shown high performance in word-level QE (Luong et al., 2013) as well as phrase-level QE (Logacheva and Specia, 2015) tasks. CRFSuite provides five optimisation algorithms: L-BFGS with L1/L2 regularization (lbfgs), SGD with L2-regularization (l2sgd), Averaged Perceptron (ap), Passive Aggressive (pa), and Adaptive Regularization of Weights (arow). Since these algorithms could perform differently in our task, we tested all of them on both baseline and extended feature sets, using the development set.

	Feature set	
	Baseline	Extended
lbfgs	0.270	0.332
l2sgd	0.238	**0.358**
ap	0.316	0.355
pa	**0.329**	**0.357**
arow	0.292	0.315

Table 1: F_1-multiplied scores of models trained on baseline and extended feature sets using different optimisation algorithms for CRFSuite.

Table 1 shows the performance of our CRF models trained with different algorithms. We can see that the extended feature set clearly outperforms the baseline for all algorithms. Passive-Aggressive scored higher for the baseline feature set and is also one of the best-performing algorithms on the extended feature set. Therefore, we used the Passive-Aggressive algorithm for our subsequent experiments and the final submission.

2.2 Data filtering

Many datasets for word-level QE suffer from the uneven distribution of labels: the “BAD” words occur much less often than those labelled as “OK”. This characteristic stems from the nature of the word-level QE task: we need to identify erroneous words in an automatically translated text, but the state-of-the-art MT systems allow producing texts of high enough quality, where only a few words are incorrect. Since for the shared task data the phrase-level labels were generated from word-level labels, we run into the same problem at the phrase level. Here the discrepancy is not so large: the “BAD” labels make for 25% of all labels in the training dataset for the phrase-level task. However, we believe it is still useful to reduce this discrepancy.

Previous experiments with word-level QE showed that the distribution of labels can be smoothed by filtering out sentences with little or no errors (Logacheva et al., 2015). Admittedly, if a sentence has no “BAD” words it lacks information about one of the classes of the problem, and thus it is less informative. We thus applied the same strategy to phrase-level QE: we ranked the

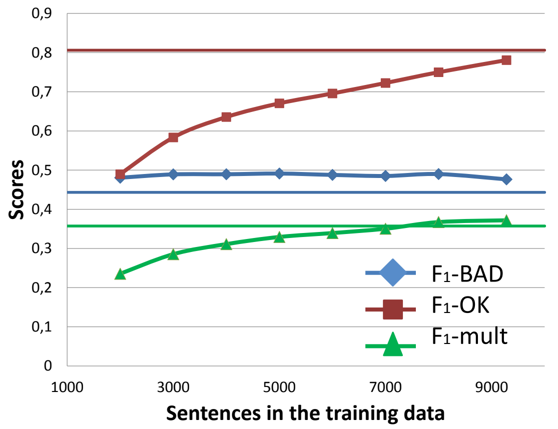

Figure 1: Performance of the phrase-level QE model with different numbers of training sentences.

training sentences by their HTER score (ratio of "BAD" words in a sentence) so that the worst sentences are closer to the top of the list, and trained our phrase-level QE model using only N top sentences from the training data (i.e. only sentences with larger number of errors).

Figure 1 shows how the scores of our phrase-level models change as we add more training data. We examine F_1-scores for both "BAD" and "OK" classes as well as their multiplication, which is the primary metric for the task (denoted as $\mathbf{F_1}$**-mult**). The flat lines denote the scores of a model that uses the entire dataset (12,000 sentences): red for F_1-OK, blue for F_1-OK, green for F_1-mult. It is clear that F_1-BAD benefits from filtering out sentences with less errors. The models with reduced data never reach the F_1-OK score of the ones which use the full dataset, but their higher F_1-BAD scores result in overall improvements in performance. The F_1-mult score reaches its maximum when the training set contains only sentences with errors (9,280 out of 12,000 sentences), although F_1-BAD score is slightly lower in this case than with a lower number of sentences. Since F_1-mult is our main metric, we use this version of the filtered dataset for the final submission.

3 Prediction-based model

Following the approach in (Specia et al., 2015), which makes use of word-level predictions at sentence level, we describe here the first attempt to using both word-level and sentence-level predictions for phrase-level QE (W&SLP4PT).

Phrase-level labels by definition depend on the quality of individual words comprising the phrase: each phrase-level label in the training data is the generalisation of word-level labels within the considered phrase. However, we argue that the quality of a phrase can also be influenced by overall quality of the sentence.

We used the following set of features based on predictions of different levels of granularity and on the phrase segmentation itself:

- **Sentence-level prediction** features:

  1. sentence score — quality prediction score assigned for the current sentence. Same feature value for all phrases in a sentence.

- **Phrase segmentation** features:

  2. phrase ratio — ratio of the length of the current phrase to the length of the sentence;
  3. phrase length — number of words in the current phrase.

- **Word-level prediction** features:

  4/5. number of words predicted as "OK"/"BAD" in the current phrase;
  6/7. number of words predicted as "OK"/"BAD" in the sentence.

Similarly to the context-based model described in Section 2, we trained our prediction-based model with the CRFSuite toolkit and the Passive-Aggressive algorithm. The phrase segmentation features are extracted from the data itself and do not need any additional information. The sentence-level score is produced by the SHEF-LIUM-NN system, a sentence-level QE system with neural network features as described in (Shah et al., 2016). The word-level prediction features are produced by the SHEF-MIME QE system (Beck et al., 2016), which uses imitation learning to predict translation quality at the word level.

4 Results

We submitted two phrase-level QE systems: the first one uses the set of baseline features enhanced with context features, the second one uses the features based on predictions made by word-level and

	F_1-BAD	F_1-OK	F_1-mult
W&SLP4PT	0.486	0.757	**0.368**
CONTEXT	0.470	0.777	**0.365**
BASELINE	0.401	0.800	0.321

Table 2: Performance of our official submissions on the test set.

sentence-level QE models, plus the phrase segmentation features. The performance of our official submissions on the test set is given in Table 2.

For the prediction-based model, we used word-level predictions from the MIME system with β=0.3. While (Beck et al., 2016) reports better performance with β = 1, we obtained slightly lower performance both on F_1-mult = 0.367 and F_1-OK = 0.739. Only F_1-BAD was better = 0.497.

Even though the two systems are very different in terms of the features they use, their performance is very similar. The prediction-based model is slightly better in terms of F_1-BAD, whereas the context-based model predicts "OK" labels more accurately. Both systems outperform the baseline.

In terms of the F_1-multiplied metric, our prediction-based and context-based systems ranked 4th and 5th (out of 10 systems) in the shared task, respectively.

4.1 Model combination

Since both our models outperform the baseline system, we also combined them after the official submission to check whether further improvements could be obtained. Surprisingly, we got the exact same prediction performance as our prediction-based model. This is because two features of our prediction-based model – the number of words predicted as "BAD"/"OK" in the current phrase – have a strong bias and do most of the job by themselves[3]. The reason of this behaviour lies in the way both the training and test data have been tagged for the phrase-level task. The labelling was adapted from the word-level labels by assigning the "BAD" tag to any phrase that contains at least one "BAD" word. Consequently, during the training against gold standards labels, our model learns to tag as "BAD" any phrase that contains at least on "BAD" word in a systematic way.

	F_1-BAD	F_1-OK	F_1-mult
W&SLP4PT	0.389	0.727	0.283
+baseline	0.454	0.767	0.349
+context	0.473	0.772	**0.366**
BASELINE	0.401	0.800	0.321

Table 3: Performance for combinations of models on the test set.

After removing the features 4 and 5 from the feature set, we retrained our prediction-based model and its new performance is given in the first row of Table 3. On its own, it performs worse than the baseline, but by successively adding the baseline and context features to it (without any data filtering), it performs as well as our official submissions in terms of F_1-BAD and F_1-multi, and gets higher F_1-OK.

5 Conclusion and future work

We presented two different approaches to phrase-level QE: one extends the baseline feature set with context information, another combines the scores of different levels of granularity to model the quality of phrases. Both performed similarly, although the prediction-based strategy is more "pessimistic" regarding the training data. Both outperformed the baseline.

In future work, we further experiments to gather a better understanding of these approaches. First, additional feature engineering can be performed: we did not check the usefulness of individual context features, nor of the additional features used in the prediction-based model. Secondly, the correspondences between labels of different granularities can be further examined: for example, it is interesting to see how the use of sentence-level and word-level predictions can influence the prediction of phrase-level scores.

Acknowledgements

This work was supported by the EXPERT (EU FP7 Marie Curie ITN No. 317471, Varvara Logacheva) and the QT21 (H2020 No. 645452, Lucia Specia, Frédéric Blain) projects.

References

Daniel Beck, Andreas Vlachos, Gustavo H. Paetzold, and Lucia Specia. 2016. SHEF-MIME: Word-level Quality Estimation Using Imitation Learning.

[3] We get the exact same scores either combining the prediction-based features with the baseline features, both the baseline and context features, or considering the number of predicted "BAD" words in the current phrase as the only feature of our model.

In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany.

Frédéric Blain, Jean Senellart, Holger Schwenk, Mirko Plitt, and Johann Roturier. 2011. Qualitative Analysis of Post-Editing for High Quality Machine Translation. In *Proceedings of the MT Summit XIII*, pages 164–171, Xiamen, China.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 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.

Simona Gandrabur and George Foster. 2003. Confidence estimation for translation prediction. In *Proceedings of Seventh Conference on Natural Language Learning*, pages 95–102, Edmonton, Canada.

Varvara Logacheva and Lucia Specia. 2015. Phrase-level quality estimation for machine translation. In *Proceedings of the 2015 International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Varvara Logacheva, Chris Hokamp, and Lucia Specia. 2015. Data enhancement and selection strategies for the word-level quality estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 330–335, Lisbon, Portugal.

Ngoc Quang Luong, Benjamin Lecouteux, and Laurent Besacier. 2013. LIG system for WMT13 QE task: Investigating the usefulness of features in word confidence estimation for MT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 386–391, Sofia, Bulgaria.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (CRFs). Available at `http://www.chokkan.org/software/crfsuite/`.

Sylvain Raybaud, David Langlois, and Kamel Smaïli. 2011. This sentence is wrong. Detecting errors in machine-translated sentences. *Machine Translation*, 25(1):1–34.

Kashif Shah, Fethi Bougares, Loic Barrault, and Lucia Specia. 2016. Shef-lium-nn: Sentence level quality estimation with neural network. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany.

Lucia Specia, G Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and Seventh International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: System Demonstrations*, pages 115–120.