Quality Estimation with Force-Decoded Attention and Cross-lingual Embeddings

Elizaveta Yankovskaya **Andre Tättar** **Mark Fishel**
Institute of Computer Science
University of Tartu, Estonia
{elizaveta.yankovskaya,andre.tattar,fishel}@ut.ee

Abstract

This paper describes the submissions of the team from the University of Tartu for the sentence-level Quality Estimation shared task of WMT18. The proposed models use features based on attention weights of a neural machine translation system and cross-lingual phrase embeddings as input features of a regression model. Two of the proposed models require only a neural machine translation system with an attention mechanism with no additional resources. Results show that combining neural networks and baseline features leads to significant improvements over the baseline features alone.

1 Introduction

Over the last several years the quality of machine translation has grown significantly. However even today most machine translation systems produce a lot of unreliable translations, with translation quality varying greatly between different input and output segments. To estimate the quality of these translations several methods have been proposed (Specia et al., 2013; Martins et al., 2017; Kim et al., 2017a,b).

In this article we propose an approach to quality estimation that is based on a regression model with different sets of features stemming from the internal parameters of a neural machine translation (NMT) system. We investigate how different input features of the regression model affect the correlation between the automatic quality estimation score and human assessment. We show that our models work for any translation output, without access to the translation system that produced the translations in question.

2 Method

The main idea of our method is to use features based on NMT attention weights and metrics based on cross-lingual embeddings as features of a regression model. In the following we explain the details of both these feature sources.

2.1 Attention Weights

The encoder-decoder NMT systems with an attention mechanism (Bahdanau et al., 2014) produce the translation output with the help of computed attention weights showing the strength of the connection between the input and output tokens. These attention weights resemble a soft alignment and their visualization often clearly indicates the translation quality that can be expected – see Figure 1 for an example of a well translated sentence.

Rikters and Fishel (2017) have shown that the attention weights can be used for confidence estimation, but only if these attention weights were computed along with translations, using the internal parameters of the NMT system producing the translations. We expand their approach to apply attention weights to any translations, regardless of whether they were produced by a data-driven, rule-based translation system or even a human translator. The same approach is used for quality estimation in (Yankovskaya and Fishel, 2018).

To get attention weights for any translation pair, we replace the decoding part of the NMT system with computing the probability of the given translation under an NMT model for that language. This way beam search and selecting the output token with the highest predicted probability is replaced with selecting the next given output token; in other words, force-decoding is done. Thus, we can get attention weights for any source/translation pair without even knowing anything about the system that produced this translation output.

To get features for a regression model we have computed the following metrics proposed by Rik-

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 816–821
Belgium, Brussels, October 31 - Novermber 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-64093

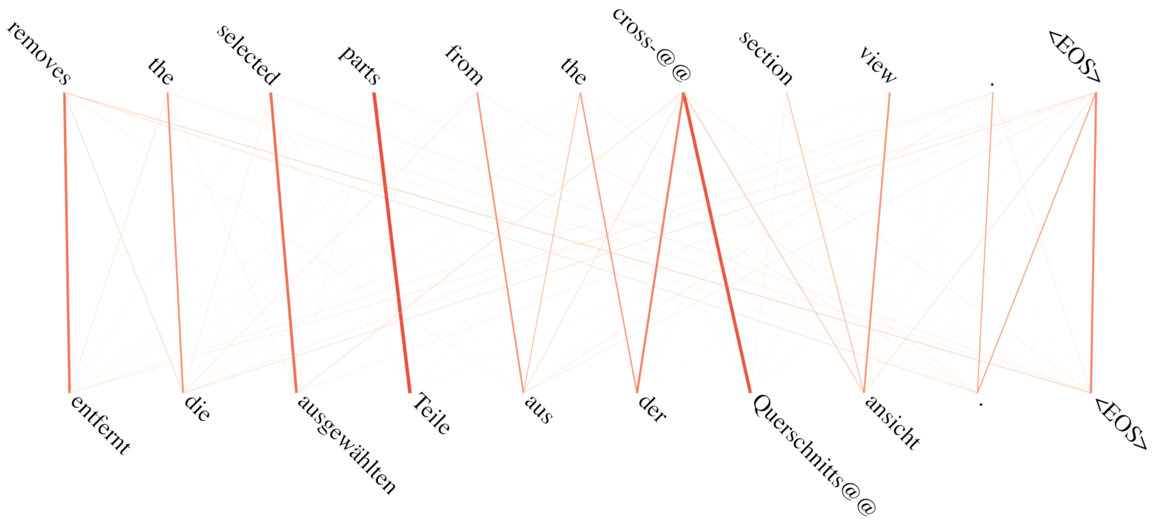

Figure 1: Attention alignment visualization of a well translated sentence from English to German. The thicker the line, the stronger the connection between the tokens (Rikters et al., 2017). It is visible from the alignment visualization alone that the quality/confidence of the translation system is high: each input/output token has a strong connection to one or at most two tokens on the other side.

ters and Fishel (2017) (see their paper for a more detailed definition):

- **Coverage Deviation Penalty (CDP)** penalizes the sum of attentions per input token, so tokens with less or too much attention get lower scores.
- **Absentmindedness Penalties (APin and APout)** compute the dispersion via the entropy of the attention distribution of input and output tokens.
- **Total** is the sum of all three metrics described above.

In addition to the metrics above we have calculated the ratio between input and output absentmindedness penalties as a small modification.

2.2 Cross-lingual Embeddings

NMT attention weights show the strength of the connection between the input and output tokens, but require running each segment pair through the NMT system. Here we try to align the input and output embeddings directly with the same aim of estimating the similarity between the input and output segments. This is done by taking the embedding-enhanced BLEU score called BLEU2VEC (Tättar and Fishel, 2017) and doing it cross-lingually.

We used three different types of embeddings to learn the cross-lingual similarity:

- **Word**-level embeddings were trained on tokenized data that consisted only of unigram words.
- **Phrase**-level embeddings were trained on data that concatenated words into phrases stochastically (Tättar and Fishel, 2017). Phrases consisted of up to three words concatenated with underscores.
- **BPE**-level embeddings use the embeddings from NMT systems that are trained on byte pair encoded data (Sennrich et al., 2015). BPE (byte-pair encoding) splits words into sub-word units in order to reduce the number of unique tokens.

The word-level and phrase-level embeddings were trained separately using monolingual corpora.Embeddings for BPE-s came from the attention-decoder translation system used in the attention weight feature extraction. These embeddings were not trained separately, so no additional training time was required for them.

After learning the monolingual embeddings, joint cross-lingual vector spaces are learned based on the monolingual ones, using the method of (Conneau et al., 2017). Cross-lingual mappings are learned between all the language pairs using MUSE[1]. In case of word-level and phrase-level mappings we used the supervised learning which

[1] A library for Multilingual Unsupervised or Su-

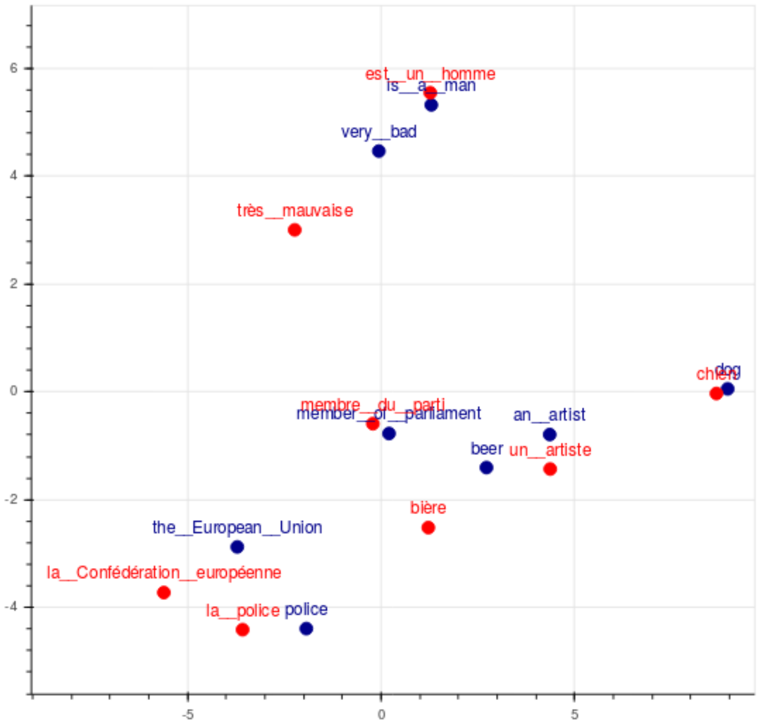

Figure 2: An examples of cross-lingual embeddings for French and English in the same vector space. On the figure, the closest neighbor was found and put on the graph. Dimensions are reduced from 300 down to the 2 first PCA components. Phrases are concatenated with two underscores. Blue means the source word/phrase and red means the nearest neighbor.

uses a seed dictionary of 1500 words for learning the mapping. For BPE embeddings we used the unsupervised cross-lingual mapping, which does not require a seed dictionary. Both methods of learning cross-lingual mappings for embedding spaces are described in (Conneau et al., 2017).

With the cross-lingual embeddings ready we compute the BLEU2VEC score:

- we find the optimal alignment between the words, n-grams or subwords of the input and output segments using beam search
- using this alignment we compute the BLEU score's (Papineni et al., 2002) n-gram precisions, giving partial credit to aligned n-gram (or word/sub-word) pairs equal to the cosine similarity of their cross-lingual embeddings

We can see examples of words/phrases after training cross-lingual embeddings in Figure 2. The nearest neighbor for a source word or phrase is visualized in the figure, which can be words or phrases in target language.

3 Experimental Settings

3.1 Data

We have applied our methods to all language pairs presented in the WMT18 shared task on sentence-level quality estimation (Specia et al., 2018): German-English, English-German, English-Latvian and English-Czech. For English-German and English-Latvian language pairs the translation output was produced by NMT and SMT systems, for other languages only SMT translations were given.

The number of sentences for each language pair and each machine translation system is shown in Table 1.

3.2 Experiments

The main goal of our experiments is to predict the normalized edit distance (HTER) (Snover et al., 2006). To estimate the quality of prediction we used the Pearson correlation coefficient.

As a regression model we used Random Forest (Ho, 1995) with a grid search algorithm for the optimization of parameters.

To get force-decoded attention weights and

pervised word Embeddings, https://github.com/facebookresearch/MUSE

	EN-DE		DE-EN		EN-CS		EN-LV	
	nmt	smt	nmt	smt	nmt	smt	nmt	smt
train	13442	26299	-	26032	-	40254	12936	11251
dev	1000	1000	-	1000	-	1000	1000	1000
test	1023	1926	-	1254	-	1920	1448	1315

Table 1: Number of sentences for each language pair and each machine translation system.

BPE embeddings for all language pairs we used NMT models trained by the University of Edinburgh (Sennrich et al., 2017) for English-German, German-English and English-Czech; for English-Latvian we used a different NMT model trained separately.

Our chosen implementation of word and phrase embeddings was FastText (Bojanowski et al., 2016) with a continuous bag-of-words (CBOW) model and the number of dimensions for embeddings was set to 300. MUSE (Conneau et al., 2017) was used for extracting cross-lingual embeddings, with default parameters. A simple beam search was implemented for finding the quality estimation BLEU2VEC score, with beam size 3.

Initial tests showed that models with features based on cross-lingual embeddings only gave a close-to-zero Pearson correlation score, therefore these were not included as standalone features into the final experiments. A combination of cross-lingual embeddings (words, phrases, BPE) demonstrated a little bit better results but they were still lower than results obtained by using a model based on the attention weights. Taking into the account these results, we ran the final experiments with the following sets of features:

- **QuEst**: a standard set of 17 black-box QuEst features (Specia et al., 2013);
- **AttW**: features based on the force-decoded attention weights: CDP, AP_{in}, AP_{out}, $total$, AP_{ratio};
- **QuEst+AttW**: a combination of QuEst and attention weights features;
- **QuEst+AttW+CrEmb3**: a combination of QuEst, attention weights and cross-lingual embeddings (phrases, words and BPE) features;
- **AttW+BPE**: a combination of attention weights and cross-lingual embeddings (BPE) features – to test a scenario of using only the parameters of an NMT system, both for the attention weights and the BPE embeddings
- **AttW+CrEmb3**: a combination of attention weights and cross-lingual embeddings (phrases, words and BPE) features.

The model with QuEst features was used as a baseline.

4 Results

The resulting Pearson coefficients for the dev and test sets for the all given language pairs are presented in Table 2. As one can see the highest values were obtained by applying the models `QuEst+AttW` or `QuEst+AttW+CrEmb3`. For English-German (NMT and SMT) and English-Latvian (SMT) language pairs the difference between these two models is negligible.

The baseline model shows the best result for all language pairs but German-English in comparison with two of our models: `AttW` and `AttW+BPE`. Although for English-Czech and English-Latvian (NMT) the difference between the baseline model and our models is small: 0.389/0.355 and 0.462/0.445. It is interesting to note that for German-English all of our proposed models showed a result that is more than twice the baseline model's result.

The main advantage of our models `AttW` and `AttW+BPE` is that they do not require additional resources like language models, n-gram frequencies, alignment probability files or even additional embedding models. In the case when the translation output is produced by an NMT system with an attention mechanism both models require attention weights or/and BPE embeddings of this NMT model. In the case when the system produced the translation is unknown one might use any NMT system with an attention mechanism.

	EN-DE				DE-EN		EN-CS		EN-LV			
	smt		nmt		smt		smt		smt		nmt	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
QuEst	0.387	0.369	0.390	0.354	0.392	0.220	0.406	0.389	0.382	0.389	0.491	0.462
AttW	0.292	0.249	0.197	0.219	0.539	0.533	0.313	0.319	0.336	0.323	0.394	0.438
AttW+ BPE	0.303	0.260	0.207	0.230	0.553	0.544	0.326	0.355	0.357	0.323	0.403	0.445
AttW+ CrEmb3	0.303	0.209	0.244	0.224	0.559	0.551	0.353	0.250	0.349	0.323	0.454	0.444
QuEst+ AttW	0.453	**0.426**	0.405	**0.373**	0.565	0.554	0.468	**0.451**	0.460	0.402	0.562	0.531
QuEst+ AttW+ CrEmb3	0.457	0.424	0.408	0.369	0.592	**0.570**	0.487	0.406	0.461	**0.404**	0.585	**0.542**

Table 2: The Pearson correlation coefficients for the dev and test sets for all language pairs.

5 Discussions

As we mentioned above, the value of the Pearson correlation coefficient for German-English language pair is much higher than the values for other language pairs. A similar result is observed for the data of the last year Quality Estimation shared task, where the resulting Pearson correlation coefficient produced by the model `AttW` was 0.302 for English-German and 0.485 for German-English. We assume that this is related to the domain of data: German-English and English-Latvian data belongs to one domain (pharmaceutical) whereas English-German and English-Czech sentences were taken from the another domain (IT). This assumption is confirmed by the fact that the values of the Pearson correlation coefficient for English-Latvian are also slightly higher than the values for other language pairs.

To investigate how the choice of the NMT system affects the Pearson correlation between an automatic prediction and human assessment, we compared the results of our NMT system and University of Edinburgh's NMT system for German-English language pair.

The resulting Pearson coefficients of two proposed models `AttW` and `QuEst+AttW` are presented in Table 3. The resulting scores differ but not significantly; although on one hand this suggests that the choice of the NMT system is not important, both of the compared NMT systems are general-domain models, equally dissimilar from both of the test data domains; a more thorough comparison is left for future explorations.

	AttW		QuEst +AttW	
	dev	test	dev	test
Edinburgh's NMT system	0.539	0.533	0.565	0.554
Our NMT system	0.560	0.562	0.594	0.584

Table 3: The Pearson coefficients of two regression models for German-English language pair. Attention weights were obtained from two different systems.

6 Conclusions

In this paper we described our submissions to the sentence-level subtask of WMT18 Quality Estimation task. We proposed several models for quality estimation of machine translation based on attention weights and embeddings. Our models do not require any additional resources, except for an NMT system and/or cross-lingual word embeddings learned from monolingual corpora. In the case when the translation output is produced by an NMT system with an attention mechanism, two of our models require only attention weights and BPE embeddings that are already created by this system.

For several language pairs the proposed models demonstrated comparable results with the baseline model. In the case of the German-English language pair all of our systems showed a much better result compared to the baseline model. Furthermore, the combination of neural networks and baseline features gave much better results than the results of the baseline model.

We plan to further experiment with the attention weights for in-domain systems and compare the scores obtained by using the internal and force-decoded attention weights.

Acknowledgments

This work was supported by the Estonian Research Council grant no. 1226.

The authors would like to thank Matīss Rikters for providing the English-Latvian NMT system used in the experiments.

References

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *CoRR*, abs/1710.04087.

Tin Kam Ho. 1995. Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on*, volume 1, pages 278–282. IEEE.

Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017a. Predictor-estimator: Neural quality estimation based on target word prediction for machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(1):3.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017b. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568.

André F. T. Martins, Fabio Kepler, and Jose Monteiro. 2017. Unbabel's participation in the wmt17 translation quality estimation shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 569–574, Copenhagen, Denmark.

Kishore Papineni, 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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Matīss Rikters and Mark Fishel. 2017. Confidence through attention. In *Proceedings of MT Summit XVI*, pages 299–311, Nagoya, Japan.

Matīss Rikters, Mark Fishel, and Ondřej Bojar. 2017. Visualizing Neural Machine Translation Attention and Confidence. volume 109, pages 1–12, Lisbon, Portugal.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, Copenhagen, Denmark.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André Martins. 2018. Findings of the wmt 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Lucia Specia, Kashif Shah, Jose GC Souza, and Trevor Cohn. 2013. Quest-a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84.

Andre Tättar and Mark Fishel. 2017. bleu2vec: the painfully familiar metric on continuous vector space steroids. In *Proceedings of the Second Conference on Machine Translation*, pages 619–622, Copenhagen, Denmark. Association for Computational Linguistics.

Elizaveta Yankovskaya and Mark Fishel. 2018. Low-resource translation quality estimation for estonian. In *Proceedings of BalticHLT: the 8th International Conference Human Language Technologies: the Baltic Perspective*, Tartu, Estonia.