Improving Machine Translation Quality Estimation with Neural Network Features

Zhiming Chen, Yiming Tan, Chenlin Zhang, Qingyu Xiang, Lilin Zhang, Maoxi Li, Mingwen Wang
School of Computer Information Engineering, Jiangxi Normal University
{qqchenzhiming, tt_yymm, zhangchenlin, qingyuxian, lilinzhang, mosesli, mwwang}
@jxnu.edu.cn

Abstract

Machine translation quality estimation is a challenging task in the WMT evaluation campaign. Feature extraction plays an important role in automatic quality estimation, and in this paper, we propose neural network features, including embedding features and cross-entropy features of source sentences and machine translations, to improve machine translation quality estimation. The sentence embedding features are extracted through global average pooling from word embedding and are trained by the word2vec toolkits, while the sentence cross-entropy features are calculated by the recurrent neural network language model. The experimental results on the development set of WMT17 machine translation quality estimation tasks show that the neural network features gain significant improvements over the baseline. Furthermore, when combining the neural network features and the baseline features, the system performance obtains further improvement.

1 Introduction

Quality estimation (QE) of machine translation estimates the quality of machine translation system outputs without human references using machine learning methods. It is often divided into two steps: first, it extracts various features from source sentences, translation outputs, and external language resources to describe the translation complexity, fluency and adequacy; and second, it predicts the quality of the translation outputs with the pre-trained machine learning model. Feature extraction is crucial to the performance of QE, and traditional methods, such as QuEst (Specia et al., 2013), extract linguistically motivated features to improve the correlation between the automatic QE and human assessment. However, extracting linguistically motivated features requires part-of-speech analysis, syntactic analysis, or semantic analysis, and these linguistic analyses relate to the target language types; this consideration limits their application in other languages. To address this problem, Shah et al. (2015a, 2016) investigated continuous space language models for sentence-level QE, and Scarton et al. (2016) proposed word embedding features for document-level QE.

Inspired by their work, we propose sentence embedding features and cross-entropy features to improve the correlation between automatic QE and human assessment and to investigate how different sentence embedding dimensions of source sentences and translation outputs, as well as the size of the training corpus, affect the system performance of QE.

2 Related work

With the great success of deep learning that has been achieved in digital image processing and automatic speech recognition, deep learning has also made tremendous breakthroughs in natural language processing, e.g., the proposition of neural network language models (Bengio et al. 2003) and neural machine translation encoder-decoder frameworks (Bahdanau et al. 2014). Therefore, many researchers have proposed deep learning approaches for the QE task. In the word-level QE task, Kreutzer et al. (2015) presented deep feed-forward neural networks to estimate the word confidence. Shah et al. (2015b) exploited word embedding as a feature to estimate whether the translation of the word is "good" or "bad" in machine translation outputs. Patel et al. (2016) applied a

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 551–555
Copenhagen, Denmark, September 711, 2017.

recurrent neural network language model to the word-level QE task.

In the sentence-level QE task， Shah et al. (2015a) extracted continuous space language model (Schwenk et al. 2007) probabilities of source sentences and machine translation outputs as features, and combined them with baseline features to improve the system performance of QE. In the WMT16 QE Task， Shah et al. (2016) further proposed forward sentence cross-entropy, sentence embedding features, and neural machine translation log-likelihood features based on their previous work. They extracted word embedding features and cross-entropy features by the continuous space language model.

In contrast to the work of Shah et al., we utilize a continuous bag-of-words model to extract the word embeddings, construct sentence embedding through global average pooling from word embeddings, and utilize a recurrent neural network language model to extract sentence cross-entropy features.

3 Neural Network Features

To overcome the problem that the traditional feature extraction method relies heavily on sentence linguistic analysis, in this paper, we exploit the latest deep learning method to extract the features of translation quality from source language sentences and its machine translations. The extracted features include sentence embedding features and sentence cross-entropy features.

3.1 The Embedding Features

Word representation learning has attracted the attention of many researchers in recent years. Especially after 2013, Mikolov et al. (2013a) released the open source word embedding learning tool: word2vec [1]. Word2vec, as a word embedding learning tool, has implemented two models: CBOW (Continuous Bag-of-words) and Skip-Gram model, inspired by the neural network language model proposed by Bengio (Bengio et al. 2003). The CBOW and Skip-Gram model remove the hidden layer processing of the neural network language model, which is time consuming, and add the optimization methods of Negative Sampling and Hierarchical Softmax (Mikolov et al. 2013b). This approach improves the accuracy of the model and accelerates the training of the model. The CBOW and Skip-Gram models are very similar. Their difference lies in that the CBOW model predicts the conditional probability of the current word by the context words, while the Skip-Gram model predicts the conditional probability of the context words by the current word. Because the training speed of the CBOW model is faster than that of the Skip-Gram model, we use the CBOW model to train the word embeddings of the source language and the target language.

The window size is set to 10, using the negative sampling optimization method. Additionally, the number of negative samples is set to 10. To accelerate the training, the sampling threshold of a high frequency word is set to 1e-5, and the iteration time is set to 15. We attempt various dimension of the word embedding, varied from 256 to 4096, to achieve best performance.

After obtaining the word embeddings of each word in the source sentence and the machine translation output, the sentence embeddings are computed by averaging them. This approach is applied to both the source sentences and the machine translation outputs. When the source sentence embedding ($\boldsymbol{V}_s$) and the machine translation output embedding ($\boldsymbol{V}_t$) are both obtained, two sentence embeddings are concatenated ($\boldsymbol{V} = [\boldsymbol{V}_s; \boldsymbol{V}_t]$) as features for the QE task.

3.2 The Cross-Entropy Features

A language model, which occupies a significant position in natural language processing, is used for the modeling of the probability distributions of the word sequences. In section 3.1, the bag-of-words model is used to obtain the word embedding features. However, the disadvantage of the bag-of-words model is that it ignores the contextual relationships between the words.

The recurrent neural network possesses sequentiality and memorability, and it performs well in sequential data modeling. Therefore, the Recurrent Neural Network Language Model (RNNLM) (Mikolov et al. 2010) was proposed and first used in automatic speech recognition and reordering of machine translations. The experimental results indicate that the RNNLM is superior to the back-off language model. Since RNNLM accounts for the word order, we extract the source language sentences and their machine translation cross-entropies as features for the QE task.

[1] https://code.google.com/p/word2vec/

The RNNLM is trained by the RNNLM toolkit[2]. The number of hidden layers is set to 100, parameter "bptt" is set to 4, and the output layer class number is set to 200. The WMT17 QE development set is used to optimize the parameters of the RNNLM. The training data is shown in section 4.1. The entropy of the WMT17 QE development set that we finally trained by the RNNLM is shown in Table 1.

WMT17 QE	language	iter	entropy
en-de	en	5	7.7549
	de	3	6.5885
de-en	en	7	5.1287
	de	12	5.8929

Table 1: The entropy of each language in the WMT17 QE development set trained by the RNNLM toolkit.

4 Experimental Results

To test the performance of the neural network features for the QE task, we conduct experiments on the development set of the WMT17 sentence-level QE task.

4.1 Experiment Set

The WMT17 sentence-level QE task contains two translation directions: English to German (en-de) and German to English (de-en). Among them, the en-de corpus concerns the IT domain, while de-en concerns the pharmaceutical domain. The training set of the en-de direction consists of 23,000 sentences; the development set consists of 1,000 sentences. The training set of the de-en direction consists of 25,000 sentences; the development set consists of 1,000 sentences. A test set of 2,000 sentences is provided for each direction. HTER (Snover et al. 2006) is provided as an estimation index for the translation quality of each training set and development set. The task of the participants is to establish a QE model to predict the HTER, with the source language sentences and their machine translations.

To train the word embedding and the RNNLM, the source side and the target side of the bilingual parallel corpus for the translation task, publicly released by the WMT evaluation campaign, are used; they include Europarl v7, Common Crawl corpus, News Commentary v8 and v11; Batch1 and Batch2, localization PO files, IT-related terms from Wikipedia[3]; WMT16 and WMT17 QE task1 corpus. The statistics of the bilingual parallel corpus are shown in Table 2, the corpus are shared for the two translation directions.

The Support Vector Regression (SVR) model is utilized for the QE. To implement the model, we use the Python machine learning toolkit: scikit-learn[4], and the radial basis function is chosen for the SVR kernel function, the grid search algorithm for parameter optimization. The metrics included Pearson's correlation coefficient (Pearson r), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Spearman's correlation coefficient (Spearman ρ) and Delta Average (DeltaAvg), which were used to evaluate the performance of the QE model. Pearson r and Spearman ρ are set as primary metrics for scoring and ranking the evaluation respectively, and higher scores mean better correlations between QE and HTER.

	English	German
Number of sentences	4.8 M	
Vocabulary size	936.0 K	1796.9 K
Number of tokens	120.8 M	115.4 M

Table 2: The statistics of the corpus size for the word embedding training and RNNLM training.

4.2 Results

We exploit SVR with different features to build the QE model. Experiments are performed on the development set of the WMT17 QE, task1. The experimental results of en-de and de-en are shown in Tables 3 and 4, respectively. The rows "Baseline" and "Word2vec" represent only used the 17 baseline features that were officially released by the evaluation campaign and only used the sentence embedding features extraction by the word2vec toolkits, while the row "Word2vec+ Baseline" represents the combination of used baseline features and sentence embedding features, and so on. The system that we finally submitted uses a combination of all of the features.

Mikolov et al. (2013c) attempt different dimensions of word embedding for the source language and the target language to achieve the best translation quality. Motivated by their work, we test the diverse dimensions of the word embedding for the source language and target language on the

[2] http://www.fit.vutbr.cz/~imikolov/rnnlm/

[3] http://www.statmt.org/wmt16/it-translation-task.html

[4] http://scikit-learn.org/stable/

Features set	Pearson r	MAE	RMSE	Spearman ρ	DeltaAvg
Baseline	0.414	13.564	18.660	0.466	8.622
Word2vec	0.502	13.104	17.734	0.520	9.455
Word2vec+Baseline	0.520	12.918	17.509	0.537	9.628
Word2vec+RNNLM	0.539	12.658	17.383	0.559	9.972
Word2vec+RNNLM+Baseline	**0.544**	**12.632**	**17.285**	**0.563**	**9.998**

Table 3：Results of the en-de direction on the development set of the WMT17 QE, task1.

Features set	Pearson r	MAE	RMSE	Spearman ρ	DeltaAvg
Baseline	0.401	13.702	18.163	0.404	6.845
Word2vec	0.504	13.290	17.171	0.456	7.984
Word2vec+RNNLM	0.554	12.382	16.492	0.496	8.732
Word2vec+Baseline	0.555	12.664	16.563	0.504	8.700
Word2vec+RNNLM+Baseline	**0.580**	**12.116**	**16.162**	**0.521**	**9.024**

Table 4: Results of the de-en direction on the development set of the WMT17 QE, task1.

	Features set	Pearson r	MAE	RMSE	Spearman ρ	DeltaAvg
WMT16	Baseline	0.399	0.132	0.175	0.438	7.42
	Our system	0.527^{3rd}	0.122	0.163	0.552^{3rd}	9.37
en-de	Baseline	0.397	0.136	0.175	0.425	7.45
	Our system	0.522^{5th}	0.126	0.163	0.545^{5th}	9.54
de-en	Baseline	0.441	0.128	0.175	0.45	6.81
	Our system	0.531^{8th}	0.130	0.167	0.52^{8th}	8.62

Table 5: The system performance on the test set of the WMT16 QE and WMT17 QE

training set. For the en-de direction, the best performance is obtained when the dimensions of the source word embedding and target word embedding are 1024 and 2048, respectively. While for de-en direction, the best performance is obtained when the dimensions of the source word embedding and target word embedding are both 2048.

Then, based on the sentence embedding features, we add the cross-entropy features extracted by the RNNLM toolkit or the baseline features. When we added the cross-entropy features, the maximum value of Pearson r increased by 9.9% on the scoring evaluation, and the maximum value of Spearman ρ increased by 7.5% on the ranking evaluation. It can be found that in the en-de direction, the result obtained by adding cross-entropy features is superior to that from adding baseline features. Finally, when we combine all of the features, the maximum value of Pearson r increases by 44.6% on the scoring task, and the maximum value of Spearman ρ increased by 29.0% on the ranking evaluation compared with the baseline.

Because the training word embedding and RNNLM require a certain size of monolingual corpus, we also investigated the effects of different corpus scales on the quality of the extracted neural network features. It was found that when the training corpus contained more than 1M sentences, the QE system performance is not reduced, and when the corpus contained less than 1M sentences, the system performance will decrease gradually as the corpus size decreases. This finding demonstrates that the training word embedding and the RNNLM are not dependent heavily on the scale of the training corpus.

Finally, Table 5 provides the results of our system and the baseline system on the test set. We take the system "Word2vec+RNNLM+Baseline" as our primary system. In WMT16 QE, the performance of our system achieves the third place. In WMT17 QE, the best result of our system achieves the fifth place. Compared with the method proposed by Shah et al (2016), we use fewer features, but achieve better result on the test set.

5 Conclusions

In this paper, we train the embedding features using the word2vec toolkit and we enrich the features with cross-entropy features extracted by RNNLM to improve the correlation between the

QE and human judgment. The experimental results show that the neural network features can significantly improve the system performance. Compared with the traditional linguistically motivated features, the extracted features of the neural network are independent of the specific language.

In the future, we will train an end-to-end pure neural network model for QE, instead of using traditional SVR methods.

Acknowledgements

This research has been funded by the Natural Science Foundation of China under Grant No.6146 2044, 6166 2031, and 6146 2045. The authors would like to extend their sincere thanks to the anonymous reviewers who provided valuable comments.

References

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

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3(2): 1137-1155.

Julia Kreutzer, Shigehiko Schamoni and Stefan Riezler. 2015. Quality Estimation from ScraTCH (QUETCH): Deep Learning for Word-level Translation Quality Estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 316-322, Lisbon, Portugal.

Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Honza Cernocky, Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of Interspeech 2010*, pages 1045-1048, Makuhari, Chiba, Japan.

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, IIya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. *arXiv preprint arXiv:1310.4546.*

Tomas Mikolov, Quoc V. Le, IIya Sutskever. 2013c. Exploiting Similarities among Languages for Machine Translation. *arXiv preprint arXiv:1309.4168.*

Raj Nath Patel, Sasikumar M. 2016. Translation Quality Estimation using Recurrent Neural Network. In *Proceedings of the First Conference on Machine Translation*, pages 819-824, Berlin, Germany.

Carolina Scarton, Daniel Beck, Kashif Shah, Karin Sim Smith and Lucia Specia. 2016. Word embeddings and discourse information for Machine Translation Quality Estimation. In *Proceedings of the First Conference on Machine Translation*, pages 831-837, Berlin, Germany.

Holger Schwenk. 2007. Continuous space language models. *Computer Speech & Language*, 21(3):492-518.

Kashif Shah, Raymond W. M. Ng, Fethi Bougares, Lucia Specia. 2015a. Investigating Continuous Space Language Models for Machine Translation Quality Estimation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1073-1078, Lisbon, Portugal.

Kashif Shah, Varvara Logacheva, Gustavo Henrique Paetzold, Frederic Blain, Daniel Beck, Fethi Bougares, Lucia Specia. 2015b. SHEF-NN: Translation Quality Estimation with Neural Networks. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 342-347, Lisbon, Portugal.

Kashif Shah, Fethi Bougares, Loïc Barrault and Lucia Specia. 2016. SHEF-LIUM-NN: Sentence-level Quality Estimation with Neural Network Features. In *Proceedings of the First Conference on Machine Translation*, pages 838-842, Berlin, Germany.

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 the 7th Conference of the Association for machine translation in the Americas*, pages 223-231, Cambridge.

Lucia Specia, Kashif Shah, Jose G. C. de Souza and Trevor Cohn. 2013. QuEst – A translation quality estimation framework. In *51st Annual Meeting of Association for Computational Linguistics: Demo Session*, pages 79-84, Sofia, Bulgaria.