The Benefit of Pseudo-Reference Translations in Quality Estimation of MT Output

Melania Duma and Wolfgang Menzel
University of Hamburg
Natural Language Systems Division
{duma, menzel}@informatik.uni-hamburg.de

Abstract

In this paper, a novel approach to Quality Estimation is introduced, which extends the method in (Duma and Menzel, 2017) by also considering pseudo-reference translations as data sources to the tree and sequence kernels used before. Two variants of the system were submitted to the sentence level WMT18 Quality Estimation Task for the English-German language pair. They have been ranked 4th and 6th out of 13 systems in the SMT track, while in the NMT track ranks 4 and 5 out of 11 submissions have been reached.

1 Introduction

The purpose of Quality Estimation (QE), as a subfield of Machine Translation (MT), is to allow the evaluation of MT output without the necessity of providing a reference translation. This would be extremely beneficial in the development cycle of a MT system, as it would permit fast and cost efficient evaluation phases. In the case of the previous Quality Estimation Shared Task (Bojar et al., 2017) together with the current campaign (Specia et al., 2018a), the purpose for the sentence level track was to predict the effort required in order to post-edit a candidate translation as measured by the Human-mediated Translation Edit Rate (HTER) (Snover et al., 2006) score.

In this paper an extension of the QE method introduced in (Duma and Menzel, 2017) is presented. Our earlier version of the metric was based on learning HTER scores using tree and sequence kernels. The kernel functions were applied not only on the source segments and the candidate translations, but also on the back-translations of the MT output into the source language. The back-translations were obtained using an online MT system.

The extension proposed in this paper uses the same input data. In addition, however, the kernel functions are defined to also consider pseudo-references as an additional source of evidence. The pseudo-references represent translations of the source segments into the target language and were obtained using the same online MT system as for the back-translation. By applying both the sequence and the tree kernels on the pseudo-references, we wanted to determine if an additional data source, even if artificially generated, would have a positive impact on our previous QE method. Throughout the rest of the paper we will refer to both the newly developed QE method as well as to its earlier version as Tree and Sequence Kernel Quality Estimation (*TSKQE*), but the variant under consideration will be marked through the use of subscripts together with superscripts.

This paper is organized as follows. In Section 2 related work is presented, focusing on kernel based QE methods. In the next section the implementation details for *TSKQE* are presented. This is followed by the evaluation setup and a discussion of the results. The paper concludes with future work ideas and final remarks.

2 Related work

The benefit of kernel functions has already been investigated in the context of Quality Estimation. In the work presented by (Hardmeier, 2011) and further expanded in (Hardmeier et al., 2012), tree kernel functions in addition to feature vectors are used to predict MT output quality. Both constituency and dependency parse trees were considered, with the Subset Tree Kernels (Collins and Duffy, 2001) being applied to the former and the Partial Tree Kernel (Moschitti, 2006a)(Moschitti, 2006b) to the latter. The evaluation results revealed that the integration of tree kernels can prove beneficial when compared to the strictly feature based QE systems.

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 776–781
Belgium, Brussels, October 31 - Novermber 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-64087

Tree kernels have also been applied in the work of (Kaljahi et al., 2014) and (Kaljahi, 2015), where a QE system is built based on Subset Tree Kernels applied for the constituency and dependency parse trees corresponding to the source and candidate translation. The kernels were also combined with a series of manually designed features, while SVM regression was used, in order to predict different automatic MT evaluation methods, like for example BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and METEOR (Denkowski and Lavie, 2014) scores.

The QE method introduced in (Duma and Menzel, 2017), *TSKQE*, is based on a linear combination between tree and sequence kernels. As a tree kernel the Partial Tree Kernel (PTK) is used, while for the sequence kernel, the Subsequence Kernel (SK) (Bunescu and Mooney, 2005) was chosen. Similarly to the previously mentioned QE methods, the kernels are applied to the source and candidate translations, but in addition also on a back-translation. The work presented in this paper builds on this method, by additionally using kernel functions for pseudo-references. Pseudo-references have been utilized before in the context of QE, but as a support for the generation of features, like for example in the work of (Soricut et al., 2012), (Shah et al., 2013) or (Scarton and Specia, 2014). In (Scarton and Specia, 2014) BLEU and TER were applied to the candidate translation and pseudo-references and their scores were used as additional features in the context of document level QE.

3 Method details

Different variants of *TSKQE* were defined in (Duma and Menzel, 2017) depending on the **level** where the kernel functions are applied (source segment, candidate translation or back-translation) and the **type** of kernel function (SK or PTK).

To indicate these distinctions we will use a notation system, where the level will be marked as a subscript attached to the *TSKQE* method name, with the possible values being *source* in case of the source segments, *basic* corresponding to both source segments and candidate translations, *back* for back-translations and *pseudo* corresponding to the newly introduced pseudo-references. In the case of the type, this will be marked as a superscript, with only two possible values, *sk* for the Sequence Kernel and *ptk* for the Partial Tree Kernel. For the variants where both kernel functions are used, the superscript will be left unfilled. Examples for this notation can be found in Tables 1 and 2.

TSKQE requires parsed input data, which was generated by means of the MATE parser (Bohnet, 2010), using English and German pre-trained models for tokenization, lemmatization, tagging and parsing itself [1]. The resulting dependency tree was further processed in order to remove the arc labels and encode all the syntactic information as tree nodes. For this, a variant of the Lexical-Centered-Tree (LCT) (Croce et al., 2011) method was applied, so that the dependency relation becomes the rightmost child of the dependency heads. For the generation of the pseudo-references and back-translations, the Google Translator Toolkit [2] was used.

The actual *TSKQE* models were built with the help of the Kernel-based Learning Platform (KeLP) library (Filice et al., 2015b) (Filice et al., 2015a), where various kernel functions and learning algorithms are integrated. For our experiments, we used the Support Vector Machine epsilon-Regression algorithm to learn the HTER scores, together with the PTK and SK implementations.

4 Evaluation

The evaluation was performed measuring the correlation between the *TSKQE* scores and the HTER gold standards. This was achieved by computing the Pearson correlation coefficient, which results in a number between -1 and 1. A score of 1 indicates that there is a perfect agreement between the two sets of scores, while a score of -1 would suggest a negative agreement. In addition to the Pearson coefficient, the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) were also calculated. For both these evaluation methods, the closer their score is to 0, the better the QE system should be considered.

The significance testing of the results was performed using the methodology presented in (Graham, 2015), which is based on pairwise testing using the Williams test (Williams, 1959). [3]

[1] All these models can be found at https://code.google.com/archive/p/matetools/downloads

[2] https://translate.google.com/toolkit

[3] The script used for computing the significance testing can be found at https://github.com/ygraham/mt-qe-eval.

	SMT			NMT		
System	**Pearson**	MAE	RMSE	**Pearson**	MAE	RMSE
$TSKQE^{sk}_{source}$	0.468	0.141	0.183	0.341	0.138	0.185
$TSKQE^{sk}_{basic}$	0.517	0.136	0.176	0.387	0.136	0.181
$TSKQE^{sk}_{basic+back}$	0.522	0.135	0.176	0.391	0.136	0.180
$TSKQE^{sk}_{basic+pseudo}$	0.512	0.135	0.177	0.407	0.135	0.179
$TSKQE^{sk}_{basic+back+pseudo}$	0.523	0.135	0.176	0.409	0.135	0.178
$TSKQE^{ptk}_{source}$	0.440	0.142	0.186	0.361	0.133	0.181
$TSKQE^{ptk}_{basic+back}$	0.517	0.136	0.176	0.376	0.136	0.181
$TSKQE^{ptk}_{basic+pseudo}$	0.507	0.136	0.178	0.391	0.135	0.180
$TSKQE^{ptk}_{basic+back+pseudo}$	0.517	0.135	0.176	0.392	0.135	0.180
$TSKQE_{basic}$	0.532	0.134	0.175	0.395	0.135	0.180
$TSKQE_{basic+back}$	**0.537**	0.133	0.174	0.400	0.136	0.180
$TSKQE_{basic+pseudo}{}^{*}$	0.523	0.134	0.176	0.414	0.134	0.178
$TSKQE_{basic+back+pseudo}{}^{*}$	0.534	0.133	0.174	**0.417**	0.135	0.178
Baseline WMT	0.359	0.147	0.195	0.264	0.129	0.184
Baseline $TSKQE^{ptk}_{basic}$	0.509	0.135	0.177	0.371	0.135	0.181

Table 1: The results of the evaluation for the different TSKQE models.

In terms of the data sets, *TSKQE* was evaluated on the English-German datasets (Specia et al., 2018b) provided by the WMT18 Quality Estimation sentence level task. In contrast to the years before, the campaign offered two tracks for this language pair: in addition to the traditional one focused on SMT systems, another one considered the evaluation of an NMT system. Both tracks used translations from the IT domain, with the data consisting of tuples made up of the source segment, the candidate translation, the reference translation and the HTER score associated to that candidate translation. For the NMT system, 13,442 tuples were made available for the training, with an additional 1,000 tuples provided for development purposes. In the case of the SMT system, the training set was larger, consisting of 26,273 instances, with the same number of 1000 tuples made available for evaluation.

We compared the performance of *TSKQE* with a weak but also with a strong baseline. The former is represented by the QE system trained only on the 17 baseline features offered by the WMT18 QE campaign organizers. The features [4] have been regularly used over the past campaigns and include, for example, the number of tokens in the source sentence or the LM probability of the target sentence. We used these baseline features not only to build the baseline system, but also integrated them into *TSKQE* by means of a Radial Basis Function (RBF) kernel. For this purpose, we applied a Z-score standardization to rescale the feature values.

For the strong baseline, we considered a variant of one of the QE systems introduced by (Hardmeier et al., 2012), based on Partial Tree Kernels applied to the source segments and candidate translations. In our notation, this would correspond to the $TSKQE^{ptk}_{basic}$ notation.

The results of the evaluation for both the NMT and the SMT tracks are presented in Table 1. We highlighted in bold the highest Pearson values. Furthermore, we marked using an asterisk the two variants which we have chosen as our submissions

[4] A list of the baseline features can be found at https://www.quest.dcs.shef.ac.uk/quest_files/features_blackbox_baseline_17

NMT Models	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.002
2	0	-	-	0.072	0.119	0.22	-	-	-	-	-	-	-	-	0
3	0	0.257	-	0.051	0.079	0.133	-	-	-	-	0.478	-	-	-	0
4	0.082	-	-	-	-	-	-	-	-	-	-	-	-	-	0
5	0.031	-	-	0.215	-	-	-	-	-	-	-	-	-	-	0
6	0.019	-	-	0.141	0.229	-	-	-	-	-	-	-	-	-	0
7	0	0.06	0.32	0.021	0.015	0.059	-	-	-	-	0.357	0.394	-	-	0
8	0	0.063	0.054	0.013	0.01	0.014	0.231	-	-	-	0.238	0.241	-	-	0
9	0	0.003	0.041	0.006	0.007	0.02	0.066	0.227	-	-	0.104	0.13	-	-	0
10	0	0.004	0.002	0.005	0.005	0.01	0.053	0.095	0.331	-	0.082	0.088	-	-	0
11	0.002	0.408	-	0.022	0.01	0.066	-	-	-	-	-	-	-	-	0
12	0.002	0.388	0.497	0.024	0.021	0.015	-	-	-	-	0.437	-	-	-	0
13	0	0.002	0.014	0.001	0	0.002	0.005	0.058	0.085	0.25	0.009	0.021	-	-	0
14	0	0.002	0.001	0.001	0	0	0.006	0.004	0.1	0.086	0.009	0.006	0.326	-	0
15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

SMT Models	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-	-	-	0.034	-	-	-	-	-	-	-	-	-	-	0
2	0	-	-	0	0.29	0.489	-	-	0.278	-	0.253	0.499	-	-	0
3	0	0.218	-	0	0.183	0.336	-	-	0.156	-	0.158	0.349	-	-	0
4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
5	0.007	-	-	0	-	-	-	-	-	-	0.404	-	-	-	0
6	0.003	-	-	0	0.152	-	-	-	0.388	-	0.174	-	-	-	0
7	0	0.006	0.128	0	0.012	0.089	-	-	0.015	0.178	0.014	0.105	0.118	-	0
8	0	0.013	0.007	0	0.009	0.023	0.252	-	0.012	0.049	0.009	0.035	0.079	0.338	0
9	0.001	-	-	0	0.418	-	-	-	-	-	0.354	-	-	-	0
10	0	0.265	0.475	0	0.182	0.33	-	-	0.038	-	0.128	0.323	-	-	0
11	0.01	-	-	0	-	-	-	-	-	-	-	-	-	-	0
12	0.002	-	-	0	0.203	0.478	-	-	0.366	-	0.065	-	-	-	0
13	0	0.262	0.472	0	0.126	0.307	-	-	0.028	0.482	0.051	0.282	-	-	0
14	0	0.057	0.098	0	0.027	0.067	0.427	-	0.003	0.025	0.007	0.036	0.032	-	0
15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

1 = $TSKQE^{sk}_{source}$	2 = $TSKQE^{sk}_{basic}$	3 = $TSKQE^{sk}_{basic+back}$
4 = $TSKQE^{ptk}_{source}$	5 = $TSKQE^{ptk}_{basic}$	6 = $TSKQE^{ptk}_{basic+back}$
7 = $TSKQE_{basic}$	8 = $TSKQE_{basic+back}$	9 = $TSKQE^{sk}_{basic+pseudo}$
10 = $TSKQE^{sk}_{basic+back+pseudo}$	11 = $TSKQE^{ptk}_{basic+pseudo}$	12 = $TSKQE^{ptk}_{basic+back+pseudo}$
13 = $TSKQE_{basic+pseudo}$	14 = $TSKQE_{basic+back+pseudo}$	15 = weak baseline

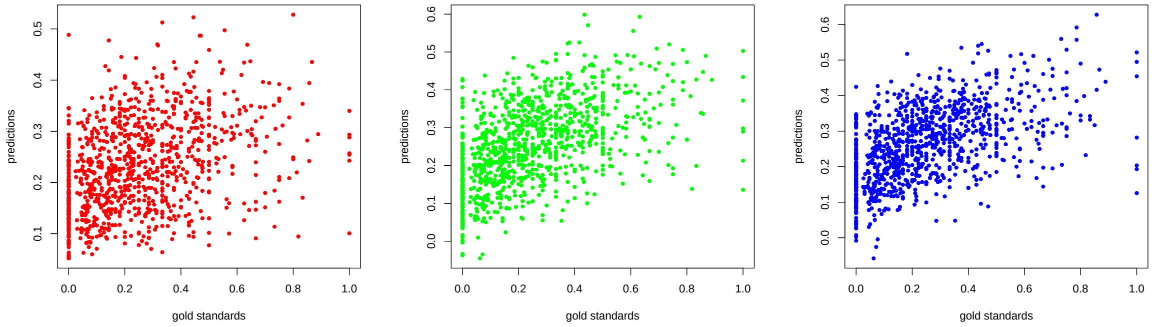
Table 2: Significance Williams test results.

to the WMT18 QE sentence level task. The results of the significance tests for two sets of *TSKQE* models are displayed in Table 2. Here, each table can be read as a matrix, where both the rows and columns correspond to the different *TSKQE* systems. The significance testing was performed only for the pairs of systems where the column model achieved a higher Pearson correlation than the row model. Otherwise, the cell was marked with a hyphen sign.

4.1 Discussion of the results

The results presented in Table 1 show that all the *TSKQE* variants outperform the weak baseline systems in terms of Pearson correlation. The same applies in the case of the strong baseline, with a few exceptions like the exclusively source based models. This result is not surprising, since the source based QE systems have access to no other input data except the source segments. The only information they receive about the candidate translation is the one contained in the baseline features.

Comparing the *TSKQE* variants based on pseudo-references with the other models, a noticeable improvement of the Pearson coefficients can be observed for the NMT system, while in the case of the SMT system the use of the pseudo-references brings no change or actually leads to a small drop in performance, which can be observed for example when comparing the *basic+pseudo* models to the *basic+back* ones. The significance tests reveal that the improvements, in the case of the NMT system, are statistically significant for the *basic+back+pseudo* models over the *basic+back* ones at a level of 0.05. In the case of the SMT system the differences between the *basic+back+pseudo* models and the *basic+back* ones are not statistically significant. In terms of



(a) SMT weak baseline model

(b) SMT strong baseline model

(c) SMT *TSKQE* model scores

Figure 1: Plots of the TSKQE and baseline model scores compared to the golden standards.

the best performing model, taking into account both MT systems, $TSKQE_{basic+back+pseudo}$, the SK and PTK based *TSKQE* variant which uses all the possible data sources, including the pseudo references, achieved on average the best correlation. These results suggest that the incorporation of the pseudo-references can be advantageous for building a high quality *TSKQE* system.

A further analysis of the results highlights the high quality of the SK based models. This is an important aspect to note, as it shows that even in the case of lower resourced language pairs, which might lack syntactic analysis tools, the SK based variants can still predict HTER scores with a comparable accuracy to the ones generated by the SK and PTK combination based models.

We also studied the degree of correlation between the predicted and the gold standard scores. Figure 1 shows the plots for the weak and the strong baseline models as well as for the $TSKQE_{basic+back+pseudo}$ model, all applied to the SMT data. [5]. Obviously, the weak baseline system encounters difficulties in predicting the HTER score as there is very little correlation between the two sets of scores. In case of the strong baseline, the predicted scores start to display a positive correlation with the gold ones, with this trend becoming even more evident in the case of the $TSKQE_{basic+back+pseudo}$ model.

5 Conclusions and future work

In this paper, we examined an extension of *TSKQE*, the sentence level QE method introduced in (Duma and Menzel, 2017). The evaluation results have not only confirmed the high quality of *TSKQE*, but they also showed that the use of pseudo-references as additional data sources for the kernel functions can be beneficial for the performance of *TSKQE*. Furthermore, the results indicate that *TSKQE* is robust against the choice of a particular MT paradigm producing comparably good results for both SMT and NMT systems.

In future work, we would like to extend the evaluation to include additional language pairs and domains. Another interesting line of research would be the use of constituency trees in addition to the dependency trees already explored to determine if these additional syntactic structures would be advantageous to the performance of *TSKQE*.

References

Bernd Bohnet. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. *The 23rd International Conference on Computational Linguistics (COLING 2010)*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Razvan Bunescu and Raymond Mooney. 2005. Subsequence Kernels for Relation Extraction. *Advances in Neural Information Processing Systems, Vol. 18: Proceedings of the 2005 Conference (NIPS)*.

Michael Collins and Nigel Duffy. 2001. Convolution Kernels for Natural Language. *Proceedings of NIPS 2001*, pages 625–632.

[5] The plots were obtained using the R language (R Core Team, 2014) and its packages

Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2011. Structured Lexical Similarity via Convolution Kernels on Dependency Trees. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1034–1046.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Melania Duma and Wolfgang Menzel. 2017. UHH Submission to the WMT17 Quality Estimation Shared Task. In *Proceedings of the Second Conference on Machine Translation*, pages 556–561.

Simone Filice, Giuseppe Castellucci, Roberto Basili, Giovanni Da San Martino, and Alessandro Moschitti. 2015a. KeLP: a Kernel-based Learning Platform in Java. *The workshop on Machine Learning Open Source Software (MLOSS): Open Ecosystems*.

Simone Filice, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2015b. KeLP: a kernel-based learning platform for Natural Language Processing. *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 19–24.

Yvette Graham. 2015. Improving Evaluation of Machine Translation Quality Estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1804–1813.

Christian Hardmeier. 2011. Improving Machine Translation Quality Prediction with Syntactic Tree Kernels. *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 233–240.

Christian Hardmeier, Joakim Nivre, and Jorg Tiedemann. 2012. Tree Kernels for Machine Translation Quality Estimation. *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 109–113.

Rasoul Kaljahi. 2015. The Role of Syntax and Semantics in Machine Translation and Quality Estimation of Machine-translated User-generated Content. *PhD Thesis*.

Rasoul Kaljahi, Jennifer Foster, Raphael Rubino, and Johann Roturier. 2014. Quality Estimation of English-French Machine Translation: A Detailed Study of the Role of Syntax. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2052–2063.

Alessandro Moschitti. 2006a. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. *Proceedings of the 17th European Conference on Machine Learning*.

Alessandro Moschitti. 2006b. Making Tree Kernels Practical for Natural Language Learning. *Proceedings of the Eleventh International Conference of the European Association for Computational Linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Carolina Scarton and Lucia Specia. 2014. Document-level translation quality estimation: exploring discourse and pseudo-references. In *The 17th Annual Conference of the European Association for Machine Translation*, pages 101–108.

Kashif Shah, Trevor Cohn, and Lucia Specia. 2013. An Investigation on the Effectiveness of Features for Translation Quality Estimation. In *Proceedings of the Machine Translation Summit*, volume 14, pages 167–174. Citeseer.

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.

Radu Soricut, Nguyen Bach, and Ziyuan Wang. 2012. The SDL Language Weaver Systems in the WMT12 Quality Estimation Shared Task. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 145–151. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André Martins. 2018a. 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, Varvara Logacheva, Frederic Blain, Ramon Fernandez, and André Martins. 2018b. WMT18 Quality Estimation Shared Task Training and Development Data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University.

Evan James Williams. 1959. *Regression analysis*, volume 14. Wiley New York.