Word embeddings and discourse information for Machine Translation Quality Estimation

Carolina Scarton, Daniel Beck, Kashif Shah, Karin Sim Smith and **Lucia Specia**
Department of Computer Science
University of Sheffield, UK
{c.scarton,debeck1,kashif.shah,kmsimsmith1,l.specia}
@sheffield.ac.uk

Abstract

In this paper we present the results of the University of Sheffield (SHEF) submissions for the WMT16 shared task on document-level Quality Estimation (Task 3). Our submission explore discourse and document-aware information and word embeddings as features, with Support Vector Regression and Gaussian Process used to train the Quality Estimation models. The use of word embeddings (combined with baseline features) and a Gaussian Process model with two kernels led to the winning submission in the shared task.

1 Introduction

The task of Quality Estimation (QE) of Machine Translation (MT) consists in predicting the quality of unseen data using Machine Learning (ML) models trained on labelled data points. Such a scenario does not require reference translations and only uses information from source and target documents. Therefore, QE is different from traditional automatic evaluation metrics (such as BLEU (Papineni et al., 2002)).

Sentence-level and word-level QE have been widely explored along the years (Callison-Burch et al., 2012; Bojar et al., 2013; Bojar et al., 2014; Bojar et al., 2015). On the other hand, document-level QE has only recently started to be addressed, with the first shared task organised last year (Bojar et al., 2015). Document-level QE is the task of predicting the quality of an entire document and is useful for *gisting* applications (mainly in cases where the user does not speak the source language) and fully automated uses of MT where post-editing is not an option.

Predicting the quality of documents is challenging: problems on all linguistic levels need be taken into account, including document-wide issues. Moreover, defining quality labels for documents is a complex task on itself, as pointed by Scarton et al. (2015b).

Little previous research has addressed this problem. Soricut and Echihabi (2010) explore pseudo-references and document-aware features for document-level ranking, using BLEU as quality label. Scarton and Specia (2014) apply pseudo-references, document-aware and discourse-aware features for document-level quality prediction, using BLEU and TER as quality scores. Last year, a paragraph-level QE shared task was organised for the first time at WMT (Bojar et al., 2015), using METEOR as quality label. Scarton (2015) explore discourse information for paragraph-level prediction. They also perform exhaustive search and find out that using only three features from the official baseline set leads to results comparable to those of the full baseline system. Biçici et al. (2015) apply referential translation machines for paragraph-level QE and obtain the best overall results in the shared task. Finally, Scarton (2015), Scarton and Specia (2015) and Scarton et al. (2015b) analyse the task of document-level QE from the perspective of defining reliable labels. They also investigate the correlation of discourse phenomena and document-lvel translation quality.

In this paper, we focus on feature engineering and the use of different ML techniques for document-level QE in the context of the WMT16 QE shared task (Task 3). We submitted two systems:

- GRAPH-BASE: counts on pronouns, connectives, Rhetorical Structure Theory (RST) and Elementary Discourse Units (EDUs) information (similar to (Scarton et al., 2015a)), plus scores from an entity graph-based model for the target documents (Sim Smith et al.,

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 831–837,
Berlin, Germany, August 11-12, 2016.

2016) were used as features. This system was trained with the Support Vector Regression (SVR) algorithm. Discourse features were combined with the official baseline features.

- EMB-BASE-GP: word embeddings from the source documents combined with the official baseline features were used to train a Gaussian Process (GP)[1] with two-kernels: one for word embeddings and one for baseline features.

In addition to the official results of our submitted systems, we experiment with other feature combinations, such as scores from graph-based entity grid coherence models extracted from source documents and word embeddings generated for target documents. In Section 2 we describe the models used in our experiments and in Section 3 we present our results.

2 Systems Description

Our submissions for the shared task explore different approaches in terms of features and modelling. We describe them in detail in what follows.

2.1 Discourse-aware system

Pronouns, Connectives, EDUs and RST features (called hereafter PCER). Following (Scarton et al., 2015a), we use information from the Charniak parser (Charniak, 2000), the Discourse Parser from Joty et al. (2013), and the Discourse Connectives Tagger from Pitler and Nenkova (2009) as features for our discourse-aware model (these features could only be extracted for English, and thus for the source documents):

- Number of pronouns;
- Number of connectives (total number and number of connectives per class);
- Number of EDU breaks;
- Number of *Nucleus* and *Satellite* relations in the RST tree;
- Number of subtrees and height of the RST tree.[2]

Latent Semantic Analysis (LSA) cohesion features (called hereafter LSA). As done in Scarton and Specia (2014), we extract the following LSA features for both source and target documents:

- Average LSA Spearman rho correlation of adjacent sentences;
- Average LSA cosine distance of adjacent sentences;
- Average LSA Spearman rho correlation of all sentences;
- Average LSA cosine distance of all sentences.

Entity graph-based features (called hereafter GRAPH-source and GRAPH-target). We use an Entity Graph Model (Sim Smith et al., 2016), which is based on the bipartite graph of Guinaudeau and Strube (2013) and tracks the occurrence of entities throughout the document, including between non-adjacent sentences. Entities are taken as all nouns occurring in the document, as recommended by (Elsner, 2011). For our experiments, a POS tagger[3] is used to identify nouns. A local coherence score is calculated directly, without any training, and represents the distribution of entities in the document. This is based on the theory that coherent texts contain salient entities. Both the sentences and entities are represented as nodes, with edges connecting the entities to the sentences they occur in. The final model score reflects the total weight of all the edges leaving a sentence, which indicates how connected such a sentence is.

We use *weighted projections* (Guinaudeau and Strube, 2013). These take the number of shared entities into account, rating the projections higher for more shared entities. We calculate the coherence score of the source documents and of the target documents and incorporate these as features.

Model We combine the described features with the official baseline ones provided by the shared task organisers and use them in an SVR with RBF kernel and hyperparameters optimised via grid search (the same as the official shared task baseline system). We use the SVR implementation available in the scikit-learn toolkit (Pedregosa et al., 2011).[4]

[1] https://sheffieldml.github.io/GPy/

[2] These features are new with respect to (Scarton et al., 2015a).

[3] http://nlp.stanford.edu/software/tagger.shtml

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

2.2 Embeddings-based system

Embedding features (called hereafter EMB-source and EMB-target). The word embeddings used in our experiments are learned with the *word2vec* tool[5] (Mikolov et al., 2013b). The tool produces word embeddings using the Distributed Skip-Gram or Continuous Bag-of-Words (CBOW) models. The models are trained using large amounts of monolingual data with a neural network architecture that aims at predicting the neighbours of a given word. Unlike standard neural network-based language models for predicting the next word given the context of preceding words, a CBOW model predicts the word in the middle given the representation of the surrounding words, while the Skip-Gram model learns word embedding representations that can be used to predict a word's context in the same sentence. As suggested by the authors, CBOW is faster and more adequate for larger datasets, so we use this model in our experiments.

The data used to train the models for English is Google's billion-word corpus[6] with the vocabulary size of 527K. The Spanish data is a combination of Europarl, News-commentary and News-crawled corpora from WMT, totalling 614M words with vocabulary size of 557K. We train 500-dimensional representations with CBOW for all words in the vocabulary of both languages. We consider a 10-word context window to either side of the target word, sub-sampling option to 1e-05, and estimate the probability of a target word with the negative sampling method, drawing 10 samples from the noise distribution.

We then extract document embeddings by averaging the word embeddings in the document (for training and test sets) from these models and use these as features. These distributed numerical representations of words as features aim at locating each word as a point in a 500-dimensional space. Given that the word embeddings were trained using a large corpus, it is expected that similar words are mapped to close points in the 500-dimensional space. Therefore, the averaged word embeddings are expected to encode information about the cohesion of the document, since it encompasses information about word usage.

Model For this submission we employ a GP over the document embeddings and the baseline features. While we did try with SVR, preliminary results using cross-validation on the training set showed better results for the GP-based model.[7] Another reason for this decision is that GP easily allows the use of kernel combinations while keeping hyperparameter optimisation efficient. We explore this idea in our submission by using a sum of two isotropic[8] kernels, one for the baseline features and another one for the embeddings.

To select the best kernel combination we perform a 10-fold cross validation scheme on the training set and select the combination which performs the best in terms of Pearson's correlation score. We also consider doing model selection by picking the model with highest likelihood on the training data, similar to the scheme used in (Preoţiuc-Pietro and Cohn, 2013). However, this resulted in a worse model when compared to the cross validation scheme. We speculate that the resulting models overfit the training data, due to its small size.

The best combination, which we use in our submission, employs two Rational Quadratic (RatQuad) kernels (Rasmussen and Williams, 2006).[9] After fixing this combination, the hyperparameters are optimised by maximising the model likelihood on the full training data.

3 Experiments and Results

Apart from word embedding features, which use external corpora for training the embeddings, our systems only use the data provided by the task organisers.

Task Our participation is in Task 3 (document-level QE) in both scoring and ranking variants. Pearson r is the official primary evaluation metric for scoring, while Spearman rho is the official primary metric for ranking.

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

[6] https://github.com/ciprian-chelba/1-billion-word-language-modeling-benchmark

[7] We also experimented with a GP for training QE models using discourse-aware features, but the results were worse than with the SVR model.

[8] An alternative would be to employ Automatic Relevance Determination (ARD), a feature weighting scheme common in GPs and other Bayesian models. However, This would add a large number of hyperparameters in our case (one per feature/dimension), making the model difficult to optimise and prone to overfitting.

[9] Besides RatQuad, we also experimented with RBF, Exponential and Matern32 kernels. RatQuad showed the best results.

Data The data of Task 3 consists of 208 documents for English-Spanish language pair, extracted from the WMT08-13 translation shared task datasets. The machine translation for each source document was randomly picked from the set of all systems that participated in the translation task. The documents were evaluated by following the two-stage post-editing method described in (Scarton et al., 2015a). In the first stage, sentences are post-edited out of context, whilst in the second stage the post-edited sentences are placed in context and any remaining mistakes are corrected. The quality scores are, then, a variation of Human-targeted Translation Edit Rate (HTER) (Snover et al., 2006) that combines results from both post-editing stages.

Baseline We use the 17 QuEst++ baseline features to train our baseline systems (Specia et al., 2015). We build a baseline system with SVR and another with GP, in order to compare our systems with comparable models.[10]

Models using discourse features and SVR The features sets we experimented with are:

- baseline + PCER + LSA + GRAPH-target + GRAPH-source;
- baseline + PCER + LSA + GRAPH-target.[11]

Our models using discourse information were trained with SVR as described in Section 2.1.

Models using word embeddings and GP The features sets we experimented with are:

- baseline + EMB-source + EMB-target;
- baseline + EMB-source;[12]
- EMB-source;

Our models using word embeddings were trained using GP as described in Section 2.2.

Model selection The best models for our submissions are selected by applying 10-fold cross-validation in the training set and choosing the model with the highest averaged Pearson r correlation. The ranks for the ranking task variant are defined by ordering the predicted values best to worst.

3.1 Results

Table 1 shows the results for our experiments with discourse-aware features and SVR for the scoring sub-task. We report results of our 10-fold cross-validation method over the training and the results on the official test set. Results in the first column (10-fold) show that both discourse feature combination lead to improvements over the baseline. However, when testing on the test set, the models do not outperform the baseline. More investigation with additional data would be necessary to draw any conclusions on the reasons behind this difference.

Results for ranking using discourse-aware features are shown in Table 2. These results are reported only for the test set. Since the ranks were obtained by using the predicted scores, we could not generate rankings when testing models in the 10-fold cross-validation experiments. For this task variant, once again the discourse-aware features do not outperform the baseline features.

Tables 3 and 4 show results for scoring and ranking, respectively, for the models using word embeddings and GP. These models outperform a baseline which was also trained with GP for all cases in our 10-fold cross-validation experiment. However, when we evaluate our models on the test set, only the combination of baseline + EMB-source or EMB-source alone are better than the baseline. In fact, our result for baseline + EMB-source in the test set is the winner of the scoring sub-task, outperforming the official baseline (0.286 in Pearson r).

For ranking (calculated only for the test set), the feature sets show a similar behaviour: the model using EMB-target does not perform better than the baseline. On the other hand, EMB-source and baseline + EMB-source outperform the baseline, with the later scoring second in the official results of the shared task. It is worth mentioning that EMB-source alone is able to outperform the baseline in both sub-tasks. This is an interesting finding since word embeddings are relatively easy to acquire and only require large raw corpora as external resources.

4 Conclusions

In this paper we presented the results of our models submitted to the WMT16 QE shared task - Task 3: document-level QE. We discussed two different models: one using discourse features and SVR and

[10] For the GP model we used RatQuad kernel.

[11] Feature combination used in our GRAPH-BASE submission.

[12] Feature combination used in our BASE-EMB-GP submission.

	10-fold	test set
baseline	0.357	**0.286**
baseline + PCER + LSA + GRAPH-target + GRAPH-source	0.423	0.284
baseline + PCER + LSA + GRAPH-target	0.424	0.256

Table 1: Pearson r correlation scores of models built with discourse-aware features and SVR.

	test set
baseline	0.354
baseline + PCER + LSA + GRAPH-target + GRAPH-source	0.282
baseline + PCER + LSA + GRAPH-target	0.285

Table 2: Spearman rho correlation scores of models built with discourse-aware features and SVR.

	10-fold	test set
baseline	0.340	0.266
baseline + EMB-source + EMB-target	0.479	0.232
baseline + EMB-source	0.493	**0.391**
EMB-source	0.481	0.319

Table 3: Pearson r correlation scores of models built with word embeddings and GP.

	test set
baseline	0.345
baseline + EMB-source + EMB-target	0.279
baseline + EMB-source	**0.393**
EMB-source	0.355

Table 4: Spearman rho correlation scores of models built with word-embeddings features and GP.

another using word-embeddings and GP.

Our results showed that using word-embeddings combined with baseline features and training a GP model with two kernels (one for the baseline features and another for the word-embeddings) achieved the most promising results, having ranked top of the scoring task variant. However, only word embeddings from the source documents were useful (word embeddings from the target documents produced worse results than the baseline). The differences between the data used to extract the embeddings for source and target can be the reason for such a result. Our hypothesis is that using bigger and more relevant data for the target language could lead to better results. Another possible reason for the low performance of target embeddings is the dimension of the vectors. Mikolov et al. (2013a) use different dimensions for source and target in order to achieve the best results. Therefore, in future work we will experiment with different dimensions. Finally, an important finding is that by using only word embeddings as features we could build a model that outperforms the baseline. Nevertheless, more investigation on the topic needs to be done in order to draw concrete conclusions.

Finally, the use of discourse-aware features did not lead to improvements over the baseline on the official test sets. Our hypothesis was that discourse information would help distinguish translations with different quality levels. However, given the tools available, most discourse-aware features (e.g. RST counts) could only be extracted for English, i.e., the source documents (perfect text). We intend to further test these features in datasets where the target language (translations) is English.

Acknowledgments

This work was supported by EXPERT (EU Marie Curie ITN No. 317471, Carolina Scarton), QT21 (H2020 No. 645452, Lucia Specia) and Cracker (H2020 No. 645357, Kashif Shah) projects and CNPq (project SwB 237999/2012-9, Daniel Beck).

References

Ergun Biçici, Qun Liu, and Andy Way. 2015. Referential Translation Machines for Predicting Translation Quality and Related Statistics. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 304–308, Lisbon, Portugal.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *The Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Ondřej Bojar, Christian Buck, Christian Federman, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Ales Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, MD.

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.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montreal, Canada.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, pages 132–139, Seattle, Washington.

Micha Elsner. 2011. *Generalizing Local Coherence Modeling*. Ph.D. thesis, Department of Computer Science, Brown University, Providence, Rhode Island.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of ACL*, pages 93–103.

Shafiq Joty, Giuseppe Carenini, Raymond T. Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 486–496, Sofia, Bulgaria.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL 2013*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *The 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Emily Pitler and Ani Nenkova. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *The Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the Fourth International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 13–16, Suntec, Singapore.

Daniel Preoţiuc-Pietro and Trevor Cohn. 2013. A temporal model of text periodicities using gaussian processes. In *2013 Conference on Empirical Methods in Natural Language Processing*, pages 977–988, Seattle, WA.

Carl E. Rasmussen and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts.

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, Dubrovnik, Croatia.

Carolina Scarton and Lucia Specia. 2015. A quantitative analysis of discourse phenomena in machine translation. *Discours - Revue de linguistique, psycholinguistique et informatique*, (16).

Carolina Scarton, Liling Tan, and Lucia Specia. 2015a. USHEF and USAAR-USHEF participation in the WMT15 QE shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 336–341, Lisbon, Portugal.

Carolina Scarton, Marcos Zampieri, Mihaela Vela, Josef van Genabith, and Lucia Specia. 2015b. Searching for Context: a Study on Document-Level Labels for Translation Quality Estimation. In *The*

18th Annual Conference of the European Association for Machine Translation, pages 121–128, Antalya, Turkey.

Carolina Scarton. 2015. Discourse and document-level information for evaluating language output tasks. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 118–125, Denver, Colorado.

Karin Sim Smith, Wilker Aziz, and Lucia Specia. 2016. Cohere: A toolkit for local coherence. In *10th International Conference on Language Resources and Evaluation*, Portorož, Slovenia. To appear.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *The Seventh biennial conference of the Association for Machine Translation in the Americas*, AMTA 2006, pages 223–231, Cambridge, MA.

Radu Soricut and Abdessamad Echihabi. 2010. TrustRank: Inducing Trust in Automatic Translations via Ranking. In *The 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level Translation Quality Prediction with QuEst++. In *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*, Beijing, China.