USHEF and USAAR-USHEF Participation in the WMT15 Quality Estimation Shared Task

Carolina Scarton[1], **Liling Tan**[2] and **Lucia Specia**[1]
[1]University of Sheffield, 211 Portobello Street, Sheffield, UK
[2]Saarland University, Campus A2.2, Saarbrücken, Germany
{c.scarton,l.specia}@sheffield.ac.uk
alvations@gmail.com

Abstract

We present the results of the USHEF and USAAR-USHEF submissions for the WMT15 shared task on document-level quality estimation. The USHEF submissions explored several document and discourse-aware features. The USAAR-USHEF submissions used an exhaustive search approach to select the best features from the official baseline. Results show slight improvements over the baseline with the use of discourse features. More interestingly, we found that a model of comparable performance can be built with only three features selected by the exhaustive search procedure.

1 Introduction

Evaluating the quality of Machine Translation (MT) systems outputs is a challenging topic. Several metrics have been proposed so far comparing the MT outputs to human translations (references) in terms of ngrams matches (such as BLEU (Papineni et al., 2002)) or error rates (such as TER (Snover et al., 2006)). However, in some scenarios, human references are not available. For example, the use of machine translation in a workflow where good enough translations are given to humans for post-editing. Another example is machine translation for *gisting* by users of online systems.

Quality Estimation (QE) approaches aim to predict the quality of MT outputs without relying on human references (Blatz et al., 2004; Specia et al., 2009). Features from source (original document) and target (MT outputs) and, when available, from the MT system are used to train supervised machine learning models (classifiers or regressors). A number of data points need to be annotated for quality (by humans or automatically) for training, using a given quality metric.

Most QE research is done at sentence level. This task has been a track at WMT shared task for the last four years (Callison-Burch et al., 2012; Bojar et al., 2013; Bojar et al., 2014). In addition to sentence level, the current edition offers for the first time a track on paragraph-level QE. Exploring quality beyond sentence level is interesting for completely automatic translation applications, i.e. without human review. For instance, consider a user looking for information on a product that has several reviews automatically translated into his/her language. This user have no knowledge about the source language. To ensure that the main message of the review is preserved, for this user the quality of each word or sentence individually is not as important as the quality of the review as a whole. Therefore, predicting the quality of the whole document (or paragraph, considering paragraph as short documents) becomes necessary.

This paper presents the University of Sheffield (USHEF) and University of Saarland (USAAR) submissions to the Task 3 of the WMT15 QE shared task: paragraph-level scoring and ranking. We submitted systems for both language pairs: English-German (EN-DE) and German-English (DE-EN).

Little previous research has been done to address document-level QE. Soricut and Echihabi (2010) proposed document-aware features in order to rank machine translated documents. Soricut and Narsale (2012) use sentence-level features and predictions to improve document-level QE. Finally, Scarton and Specia (2014) and Scarton (2015) introduced discourse-aware features, which are combined with baseline features adapted from sentence-level work, in order to predict the quality of full documents. Previous work led to some improvements over the baselines used. However, several problems remain to be addressed for improving document-level QE, such as the choice of quality label, as discussed by Scarton et

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 336–341,
Lisboa, Portugal, 17-18 September 2015.

al. (2015).

Our approach focuses on extracting various features and building models with different combination of these features. Two feature selection approaches are considered. The first one is based on Random Forests and backward feature selection. The second performs an exhaustive search on the entire feature space. Features are either based on previous work for sentence-level QE (e.g. number of tokens in the target document) or are discourse-aware (e.g. lexical repetition counts).

2 Document-level features

Along with the official baseline features, we use two different sets of features. The first set contains document-aware features, based on QuEst features for sentence-level QE (Specia et al., 2013; Specia et al., 2015). The second set are features that encompass discourse information, following previous work of Scarton and Specia (2014) and Scarton (2015).

2.1 Document-aware features

The 17 baseline features made available by the organisers are the same baseline features used for sentence-level QE, adapted for document-level.[1] However, as part of the QuEst framework, other sentence-level features can be easily adapted for document-level QE. Our complete set of document-aware features include:

- ratio of number of tokens in source and target (and in target and source)
- absolute difference between number tokens in source and target, normalised by source length
- language model (LM) perplexity of source/target document (with and without end of sentence marker)
- average number of translations per source word in the document (threshold: prob $>0.01/0.05/0.1/0.2/0.5$)
- average number of translations per source word in the document (threshold: prob $>0.01/0.05/0.1/0.2/0.5$) weighted by the frequency/inverse frequency of each word in the source corpus
- average unigram/bigram/trigram frequency in quartile 1/2/3/4 of frequency in the corpus of the source language
- percentage of distinct unigrams/bigrams/trigrams seen in a corpus of the source language (in all quartiles)
- average word frequency: on average, each type (unigram) in a source document appears n times in the corpus (in all quartiles)
- percentage of punctuation marks in source/target document
- percentage of content words in the source/target document
- ratio of percentage of content words in the source and target
- LM log probability of POS of the source/target document
- percentage of nouns in the source/target document
- percentage of verbs in the source/target document
- ratio of percentage of nouns in the source and target documents
- ratio of percentage of verbs in the source and target documents
- ratio of percentage of pronouns in the source and target documents
- number of dependencies with aligned constituents normalised by the total number of dependencies (maximum between source and target)
- number of sentences (source and target should be the same).

2.2 Discourse-aware features

Discourse is a linguistic phenomenon that happens document-wide and should be considered for document-level evaluation purposes. We considered the discourse-aware features presented in Scarton and Specia (2014), which are already implemented in the QuEst framework (called herein as discourse repetition features):

- word/lemma/noun repetition in the source/target document
- ratio of word/lemma/noun repetition between source and target documents.

Other discourse features were also explored (following the work of Scarton (2015)):

- number of pronouns in the source/target document
- number of discourse connectives in the source/target document
- number of pronouns of each type according to Pitler and Nenkova (2009)'s classification:

[1] http://www.quest.dcs.shef.ac.uk/quest_files/features_blackbox_baseline_17

Expansion, *Temporal*, *Contingency*, *Comparison* and *Non-discourse*
- number of EDU (elementary discourse units) breaks in the source (target) document
- number of RST (Rhetorical Structure Theory) *Nucleus* relations in the source/target document
- number of RST *Satellite* relations in the source/target document.

In order to extract the last set of features we use existing NLP tools: For identifying pronouns, we use the output of Charniak's parser (Charniak, 2000) (we count the PRP tags). Discourse connectives are automatically extracted by the parser of Pitler and Nenkova (2009). RST trees and EDUs are extracted by the discourse parser and discourse segmenter of Joty et al. (2013).

3 Experiments and results

Our systems use only the data provided by the task organisers. For features that require corpora or resources, only those provided by the organisers were used.

Tasks we participate in Task 3 (paragraph-level QE) in both subtasks, scoring and ranking. The evaluation for the scoring task was done using Mean Absolute Error (MAE) and the evaluation for the ranking task was done by DeltaAvg (official metrics of the competition).

Data the official data of Task 3 - WMT15 QE shared task consist of 1215 paragraphs for EN-DE and DE-EN, extracted from the corpora of WMT13 machine translation shared task (Bojar et al., 2013). For training, 800 paragraphs were used and, for test, 415 paragraphs were considered. METEOR (Banerjee and Lavie, 2005) was used as quality labels.

Feature combination we experimented with different feature sets:
- baseline (17 baseline features only)
- baseline + discourse repetition features[2]
- baseline + document-aware features
- baseline + discourse-aware features
- all features.

Backward feature selection[3] in order to perform feature selection, we used the Random Forest algorithm, as implemented in the scikit-learn toolkit (Pedregosa et al., 2011), to rank the features. Once this feature ranking is produced, we apply a backward feature selection approach. Starting with the features with lower positition in the rank, the method consists in consistently eliminate features, aiming to obtain a feature set that better fit the predictions.

For both EN-DE and DE-EN, 38 features were selected. The set of features selected for both languages is:
- LM probability of source document
- LM perplexity of source document
- average trigram frequency in quartile 1/2/3/4 of frequency in a corpus of the source language
- percentage of distinct trigrams seen in a corpus of the source language (in all quartiles)
- ratio of percentage of pronouns in the source and target documents
- average number of translations per source word in the document (threshold: prob >0.1)
- average number of translations per source word in the document (threshold: prob >0.1) weighted by the frequency of each word in the source corpus
- noun/word/lemma repetition in the source document
- noun/lemma repetition in the target document
- ratio of noun/lemma/word repetition between source and target
- number of punctuation marks in the target document
- number of sentences in the source document
- number of connectives in the source document
- number of connectives in the *Expansion*/*Contingency*/*Comparison*/*Temporal*/*Non-discourse* class
- number of pronouns
- number of EDU breaks in the source document
- number of RST *Nucleus*/*Satellite* relations in the source document.

Features selected for EN-DE only:
- LM probability of target document
- LM perplexity of target document (with and without sentence markers)
- type/token ration
- average number of translations per source word in the document (threshold: prob $>0.2/0.5$)

[2]Official submission of USHEF team for EN-DE
[3]Official submission of USHEF team for DE-EN

- number of punctuation marks in the source document.

Features selected for DE-EN only:

- average source token length
- LM perplexity of source document (without sentence markers)
- average bigram frequency in quartile 1/2/3/4 of frequency in a corpus of the source language
- average number of translations per source word in the document (threshold: prob >0.01)
- average number of translations per source word in the document (threshold: prob >0.2) weighted by the inverse frequency of each word in the source corpus
- ratio of percentage of verbs in the source and target.

Exhaustive search[4] We investigate the efficacy of the baseline features by learning one Bayesian Ridge classifier for each feature and evaluating the classifiers based on MAE.

To examine the best set of features among the baseline features, we implemented an exhaustive feature selection search by enumerating all possible feature combinations. Given *n* number of features, *S*, there are 2^n-1 number of possible feature combinations since a *k*-combination of a set forms a subset of *k* distinct elements of *S*. The set of *n* elements, the number of *k*-combination is equal to the binomial coefficient:

$$(nk) = \frac{n(n-1) \quad ... \quad (n-k+1)}{k(k-1)...1} \quad (1)$$

And the sum of all possible *k*-combinations:

$$\sum_{0 \leq k \leq n} (nk) = \quad 2^n - 1 \quad (2)$$

We note that the exhaustive search for feature selection is only possible in low feature space but from the results above, it is possible to approximate the best feature combination by using the N-best performing features when the classifier is trained solely on each of the feature.

For both languages, the exhaustive search selected three features only. For EN-DE:

- average source token length
- percentage of unigrams in quartile 4 of frequency of source words in a corpus of the source language
- percentage of trigrams in quartile 4 of frequency of source words in a corpus of the source language.

For DE-EN:

- type/token ratio
- percentage of unigrams in quartile 1 of frequency of source words in a corpus of the source language
- percentage of trigrams in quartile 1 of frequency of source words in a corpus of the source language.

Machine learning algorithms for the feature combination experiments (with backward feature selection) we used the SVR implementation in the scikit-learn toolkit with parameters optimised via grid search.

3.1 Results

Table 1 shows the results of all experiments, for both language directions (EN-DE and DE-EN) and for scoring (MAE) and ranking (DeltaAvg) subtasks.[5]

For EN-DE, BFF showed the best result for scoring, and Baseline + discourse repetition showed the best result for ranking. For DE-EN, Backward feature selection showed the best results for both scoring and ranking (although BFF showed similar results for scoring).

However, no statistically significant difference was found between the systems. This means that the use of sophisticated discourse-aware features did not lead to improvements, with a simple combination of three features from the baseline set able to produce similar results. The reason for these results is most likely connected to the data. We expect the discourse-aware features to work better with documents, since they naturally contain discourse phenomena. However, the data of the shared task consists of short paragraphs, many with only one sentence only. In this case, discourse-aware features are less effective.

BFF systems investigate the efficacy of the baseline features by learning one Bayesian Ridge classifier for each feature and evaluating the classifiers based on the Mean Average Error (MAE).

[4] Official submission of USAAR-USHEF team for both language pairs - called BFF

[5] All experiments were applied to the official test set of Task 3. In order to improve readability, results for MAE and DeltaAvg were multiplied by 100.

	English-German		German-English	
Experiment	MAE ↓	DeltaAvg ↑	MAE ↓	DeltaAvg ↑
Baseline	10.05	1.6	7.35	0.59
Baseline + discourse repetition	9.55	**4.55**	6.60	1.02
Baseline + discourse-aware	9.67	4.38	7.06	1.31
Baseline + document-aware	9.57	4.55	7.68	0.37
All	9.58	4.47	6.63	0.91
Backward feature selection	10.00	3.40	**6.54**	**1.55**
BFF	**9.37**	3.98	6.56	0.4

Table 1: Results of all combinations of features

No.	Baseline Feature	MAE (DE-EN)	MAE (EN-DE)
1	number of tokens in the source document	7.21	11.69
2	number of tokens in the target document	7.31	10.81
3	average source token length	7.02	9.97
4	LM probability of source document	7.32	11.39
5	LM probability of target document	7.93	11.79
6	type/token ratio	**6.61**	9.95
7	average number of translations per source word in the document (threshold: prob >0.2)	7.49	10.70
8	average number of translations per source word in the document (threshold: prob >0.01) weighted by the inverse frequency of each word in the source corpus	6.67	**9.84**
9	percentage of unigrams in quartile 1 of frequency in a corpus of the source language	**6.61**	10.11
10	percentage of unigrams in quartile 4 of frequency in a corpus of the source language	6.72	**9.81**
11	percentage of bigrams in quartile 1 of frequency in a corpus of the source language	6.62	10.00
12	percentage of bigrams in quartile 4 of frequency in a corpus of the source language	6.64	10.05
13	percentage of trigrams in quartile 1 of frequency in a corpus of the source language	**6.59**	10.01
14	percentage of trigrams in quartile 4 of frequency in a corpus of the source language	6.62	9.97
15	percentage of unigrams in the source document seen in a corpus (SMT training corpus)	6.76	**9.75**
16	number of punctuation marks in source document	6.71	10.10
17	number of punctuation marks in target document	6.72	10.00

Table 2: MAE of classifiers trained with one baseline feature - the top three features are shown in bold

Table 2 shows the MAE of these classifiers.

We note that the exhaustive feature selection search is only possible in low feature spaces. However from the results above it is possible to approximate the best feature combination by using the N-best performing features when the classifier is trained solely on each of the feature. Unsurprisingly, the best feature set for DE-EN corresponds to the top three features that are most effective individually (when classifiers were built for these features individually). In the reverse direction (EN-DE), the best feature combination corresponds to the top 6 features that are most effective individually. The classifier trained on the top 3 features (8, 10, 15) for EN-DE yielded an MAE of 9.72.

MAE (DE-EN)	Feature Set	MAE (EN-DE)	Feature Set
6.56	(6, 9, 13)	**9.37**	(3, 10, 14)
6.57	(6, 13)	9.42	(3, 10, 13, 14)
6.59	(13)	9.43	(3, 8, 10, 11, 13, 14)
6.60	(9, 11, 13)	9.43	(3, 10, 11, 13, 14)
6.60	(9, 13, 17)	9.45	(3, 8, 10, 11, 13)

Table 3: Top five feature combinations with the lowest MAE

4 Conclusions

In this paper we presented the submissions from the USHEF and USAAR-USHEF teams for WMT15 QE shared task. We experimented with several feature combinations and used two types of feature selection methods: backward based on Random Forests and exhaustive search.

With the exhaustive search results, we showed that it is possible to build good quality regressors that outperform the baseline.

Acknowledgements

This work was supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement n° 317471.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence Estimation for Machine Translation. In *COLING 2014*, pages 315–321, Geneva, Switzerland.

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 *WMT13*, pages 1–44, Sofia, Bulgaria.

Ondrej 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 *WMT14*, pages 12–58, Baltimore, MD.

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 *WMT12*, pages 10–51, Montréal, Canada.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *NAACL 2000*, pages 132–139, Seattle, Washington.

Shafiq Joty, Giuseppe Carenini, Raymond T. Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *ACL 2013*, pages 486–496, Sofia, Bulgaria.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002*, pages 311–318, Philadelphia, PA.

Fabian Pedregosa, Gael Varoquaux, Alexander 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 Édouard 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 *ACL-IJCNLP 2009*, pages 13–16, Suntec, Singapore.

Carolina Scarton and Lucia Specia. 2014. Document-level translation quality estimation: exploring discourse and pseudo-references. In *EAMT 2014*, pages 101–108, Dubrovnik, Croatia.

Carolina Scarton, Marcos Zampieri, Mihaela Vela, Josef van Genabith, and Lucia Specia. 2015. Searching for Context: a Study on Document-Level Labels for Translation Quality Estimation. In *EAMT 2015*, pages 121–128, Antalya, Turkey.

Carolina Scarton. 2015. Discourse and document-level information for evaluating language output tasks. In *NAACL-SRW 2015*, pages 118–125, Denver, Colorado.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *AMTA 2006*, pages 223–231, Cambridge, MA.

Radu Soricut and Abdessamad Echihabi. 2010. TrustRank: Inducing Trust in Automatic Translations via Ranking. In *ACL 2010*, pages 612–621, Uppsala, Sweden.

Radu Soricut and Sushant Narsale. 2012. Combining Quality Prediction and System Selection for Improved Automatic Translation Output. In *Proceedings of WMT 2012*, pages 163–170, Montréal, Canada.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *EACL 2009*, pages 28–37, Barcelona, Spain.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. Quest - a translation quality estimation framework. In *ACL 2013*, pages 79–84, Sofia, Bulgaria.

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