Bilingual Embeddings and Word Alignments for Translation Quality Estimation

Amal Abdelsalam[*], **Ondřej Bojar**[**], **Samhaa El-Beltagy**[*]
[*]Nile University in Egypt, Center of Informatics Science, Text Mining Research Group
am.mahmoud@nu.edu.eg, samhaa@computer.org
[**]Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
bojar@ufal.mff.cuni.cz

Abstract

This paper describes our submission UFAL_MULTIVEC to the WMT16 Quality Estimation Shared Task, for English-German sentence-level post-editing effort prediction and ranking. Our approach exploits the power of bilingual distributed representations, word alignments and also manual post-edits to boost the performance of the baseline QuEst++ set of features. Our model outperforms the baseline, as well as the winning system in WMT15, Referential Translation Machines (RTM), in both scoring and ranking sub-tasks.

1 Introduction

Recently, the task of quality estimation (QE) for machine translation (MT) output attracted interest among researchers in the machine translation community. QE systems play an important role in improving post-editing efficiency (in terms of the time and effort) in different ways, e.g. by filtering out low quality translations to avoid spending time post-editing them, or by providing end-users with an estimate on how good or bad the translation is.

In 2012, WMT established the first sentence-level quality estimation shared task (Callison-Burch et al., 2012). Since then, new sub-tasks, language pairs and datasets in different domains were introduced every year (Bojar et al., 2013, 2014, 2015). In contrast to automatic evaluation (the "metrics task"), QE task aims to develop systems that provide predictions on the quality of machine translated text without access to reference translations (Blatz et al., 2004; Specia et al., 2009).

Sentence-level QE is the most popular track in the WMT QE shared task, due to its presence in all editions of the task since the beginning. Many features have been explored by participating systems, including lexical, syntactic, semantic, embedding-based features (Shah et al., 2015), as well as features dependent on any details the particular MT systems may provide (Soricut et al., 2012; Camargo de Souza et al., 2013). In our model, we try to exploit the power of bilingual distributed representations combined with word alignment information to boost the performance of translation quality estimation. For this purpose, we use the implementation provided by the Multivec tool (Bérard et al., 2016) for the bilingual distributed representation model, described by Luong et al. (2015) and the GIZA++ word alignment model (Och and Ney, 2003).

The rest of this paper is organized as follows. In Sections 2 and 3, we give an overview of the bilingual distributional model and word alignment for our purposes. Section 4 gives a detailed description of our feature set, including the features derived from manual post-edits of other sentences. Section 5 describes the datasets and resources we used to build our model. Section 6 discusses the experiments conducted and the official results. The final Section 7 concludes the paper.

2 Bilingual Distributed Representations

Word embeddings have shown a great potential in tackling various NLP tasks recently, including multilingual tasks. However, there is a major problem with using word embeddings in a multilingual setting because models are trained independently for each of the languages and the resulting

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 764–771,
Berlin, Germany, August 11-12, 2016.

representations can use the vector space very differently. Therefore, measuring similarity between words in different languages will be difficult because even similar words would likely have very different representations. Much research work has been conducted to address this problem. According to Luong et al. (2015), the approaches developed to learn bilingual models fall into three categories:

Bilingual Mapping , where word representations are trained for each language independently and a linear mapping is then learned to transform representations from one language to another (Mikolov et al., 2013a).

Monolingual Adaptation relies on pre-trained embeddings of the source language when learning target representations. A bilingual constraint (such as unsupervised word alignments derived from a parallel corpus; Zou et al., 2013) ensures that semantically similar words across languages end up with embeddings similar in the learned vector space.

Bilingual Training aims to jointly learn representations for both languages using a parallel corpus. There were attempts to jointly learn representations without relying on word alignments (Gouws et al., 2014; Hermann and Blunsom, 2014; Chandar A P et al., 2014) but the BiSkip model introduced by Luong et al. (2015) clearly benefits from word alignments and outperforms other approaches in bilingual tasks such as cross-lingual document classification.

In our submission, we use the BiSkip bilingual model, belonging to the Bilingual Training category, to measure the similarity between the source and target sentences using their compositional vector representations, where the term *compositional* indicates that the vector for the sentence is a simple sum of the vectors of all words.

BiSkip model adapts Mikolov et al. (2013b) skipgram model for the bilingual case. The joint representations are learnt using Algorithm 1 to the following objective:

$$\alpha(Mono_1 + Mono_2) + \beta Bi \qquad (1)$$

where $Mono_1$ and $Mono_2$ are the monolingual representations of each language, Bi is used tie the two monolingual spaces, and the hyperparameters α and β are used to balance the influence of the monolingual components over the bilingual one.

Data: Word-Aligned Parallel Corpus
Output: BiSkip Vector Representation
for *source-target sentence pair* **do**
  for *a(w_s,w_t) ∈ set of alignment links* **do**
    Predict neighbors of w_s;
    Predict neighbors of w_t;
    Use w_s to predict neighbors of w_t;
    Use w_t to predict neighbors of w_s ;
  end
end

Algorithm 1: BiSkip learning algorithm by Luong et al. (2015)

3 Word Alignments

For cross-lingual semantic similarity, a word alignment model is an important component. According to the evaluation of the semantic textual similarity task in SemEval 2015, the best performing systems in both the English and Spanish subtasks relied mainly on word alignment techniques (Sultan et al., 2015; Hänig et al., 2015). Inspired by these results, we add features based on word alignment to the QE system.

According to Specia et al. (2015), alignment-based features are used for word-level QE only and there is no alignment-based features included in the baseline feature set for sentence-level QE.

We use GIZA++ (Och and Ney, 2003) to obtain the alignments. By default, GIZA++ alignments are not symmetric. We symmetrize them by taking the intersection of the two directions, leading to high-precision alignments. For pre-processing, we lowercase and stem words (naively taking just the first four letters) on both sides of the input.

Some of our features rely on the alignments of our training data (the ITcorpus and the training part of the QEcorpus, see Section 5 below) and some need alignments between the source and the evaluated translation candidates (the development and test part of the QEcorpus). We thus use two sets of alignments:

Run-1 obtained by aligning only the ITcorpus.

Run-2 obtained by aligning the ITcorpus concatenated with the QEcorpus.

4 Features

This section describes the different types of features we use in our QE system. We extend the set of baseline features (Section 4.1) with features based on bilingual embeddings (Section 4.2), word alignment (Section 4.3) and also n-grams seen in a collection of manually post-edited texts (Section 4.4).

4.1 QuEst++ Baseline Features

A set of 17 system-independent features was developed by Specia et al. (2013) to set the baseline system for QE tasks. The features set is extracted using QuEst++[1] (Specia et al., 2015), an open source implementation of the baseline for quality estimation for different granularities (sentence, word, and document level QE).

QuEst++ extracts features from either or both the source and target sides (i.e. the source sentence and the candidate translation), and also language model features relying on large monolingual data.

4.2 Bilingual Embedding Features (BE)

In our submission, we use three features derived from bilingual embeddings:

SentSim simply takes the value of cosine similarity between the source and target sentences in the bilingual compositional vector space.

WordSim uses the bilingual vector model and also word-alignment links. We take the average value of cosine similarity between source words and their aligned counterparts in the target sentence. The alignment links between the source and target are established automatically. Specifically, we use Run-2 alignments as defined in Section 3.

NounSim is similar to WordSim, but instead of taking all alignment links, we compute the average cosine similarity of only the links where the source (English) word is a noun. The POS tags were produced by Stanford POS Tagger (Toutanova et al., 2003).

4.3 Alignment-Based Features

We propose several features based on automatic word alignments as obtained in Section 3.

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

4.3.1 Alignment Quality Score

We assume that a good translation aligns well word-by-word with the source. While this need not be the case for human translations, it usually holds for machine-translated text. To assess the translation quality of a segment, we thus take an alignment quality score.

In our submission, alignment quality scores are inspired by components of the conditional probability $P(t_1...t_l|s_1...s_m, a_1...a_m)$, where s_i denotes the source words, t_j denotes the target words and a_i are the alignment links for each source word to the target (unambiguous, due to the intersection). We define the score as:

$$\text{score} = \sum_{i=1}^{m} P(t_j|s_{i_{a_i}}) \tag{2}$$

$$P(t_j|s_{i_{a_i}}) = \frac{c(s_i, t_j)}{c(s_i)} \tag{3}$$

The score is a simple sum of lexical translation probabilities (longer sentences with more aligned words thus get a higher score) and the lexical translation probabilities $P(t|s)$ are estimated from the count $c(s, t)$ how often the source s and target t words were aligned in our word-aligned corpus.

The formulas resemble IBM Model 1 (Brown et al., 1993), but the counts used to compute our probability estimates are based on the whole sequence of GIZA++ models and after the heuristic symmetrization.

Run-1 alignment (see Section 3) is used in this step to avoid unreliable alignments that could be produced from aligning the poor machine translation examples in the QE datasets.

4.3.2 POS Alignment Features

Two more alignment-based features were introduced to estimate translation quality of each source-target sentence pair with the help of their POS tags. In our experiments, we restrict the range of POS tags used to produce our features to only nouns, verbs, adverbs, and adjectives. The POS tags for both English and German come again from Stanford POS Tagger.

The two introduced features are:

Number of correctly matched tags represents the number of source words that are aligned to target words with the same POS tag.

Number of wrongly matched tags represents the number of source words that are aligned to target words with a different POS tag.

Since the alignments are needed for the source and candidate translation, they come from Run-2.

4.4 Post-Edited N-grams

As mentioned earlier, in quality estimation, there is no access to reference translation. However, the QE task organizers provided the participants with training data (called "QEcorpus" in Section 5) consisting of 12k training segments and 1k development segments machine-translated and manually post-edited. To benefit from this valuable resource, we introduce another set of features representing the most frequent bigrams in translation text that were changed through the post-editing.

The list of bigrams was extracted on the basis of GIZA++ alignment, preprocessing tokens and symmetrizing the two directions the same way as in Section 3. We extract all word-aligned bigrams occurring more than 10 times in the training and development 13k sentences, greatly reducing the number of bigrams to a few dozens of most general ones. Each of the bigrams serves as an independent boolean feature in the model.

Although lowercasing seems to be more helpful during the alignment, we avoid it during the actual bigram extraction since case changes are mostly rightful and important post-edits when translating into German. On the other hand, the order in which the words and their alignments are occurred in the text is checked to be reserved (e.g. bigrams with the second target word positioned before the first target word are excluded).

Table 1 summarizes the number of extracted bigrams. Lowercased n-grams would be more general so more would survive the thresholding, but we opt to use the cased n-grams.

Lowercasing	Extracted Bigrams	Thresholded (>10)
On	71294	80
Off	73313	74

Table 1: Extracted Bigrams Numbers

Having that the 1k development segments are used to extract the N-grams features, we report the performance of the N-grams features on the 2k testing segments only.

5 Data

Our experiments use the following corpora:

QEcorpus (our name) denotes the English-German corpus released by the WMT16 QE task organizers. It is the first time when this language pair appears in the segment-level QE. QEcorpus consist of 15k source sentences in the IT domain, divided into 12k training, 1k development and 2k testing segments. Source sentences are provided with their machine translations, post-editions and HTER (Snover et al., 2006) as post-editing effort scores.

ITcorpus denotes the parallel English-German domain-specific resources made available for the WMT IT-Domain Translation Task[2]. ITcorpus consist of 452546 parallel sentences assembled from different resources, see Table 2.

Data Source	Sent. Pairs
Cross-lingual help-desk service	2000
IT related terms from Wikipedia	23134
Technical Documentation (LibreOffice, Chromium, Ubuntu)	427412

Table 2: ITcorpus sources

ComparableNews is a pair of monolingual corpora, namely the English and German versions of the News Crawl monolingual corpus (only the year 2015) compiled from various online news publications for the WMT News Translation Task[3]. The corpus consists of 3.2 GB of English text with 27.2 million sentences and 5.5 GB of German text with 51.3 million sentences. The vocabulary size for this corpus is 1,774,792 English and 5,817,655 German words (excluding numbers and punctuation).

As pre-processing, the corpus used in each setup is first cleaned from hyperlinks and then tokenized using Moses tokenizer[4].

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

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

[4] http://www.statmt.org/moses/

Features	Pearson's r	MAE	RMSE	Spearman's ρ
Baseline (QuEst++)	0.350	14.515	19.332	0.395
Baseline + AlignQualityScore	0.365	14.434	19.216	0.407
Baseline + POSAlignment	0.349	14.560	19.347	0.388
Baseline + BE_SentSim	0.353	14.487	19.303	0.399
Baseline + BE_WordSim	0.349	14.518	19.337	0.395
Baseline + BE_NounSim	0.353	14.487	19.311	0.399
All Features	**0.374**	**14.362**	**19.144**	**0.412**

Table 3: Evaluation of the introduced features using WMT16 Sentence-Level QE Development set

Features	Pearson's r	MAE	RMSE	Spearman's ρ
Baseline (QuEst++)	0.347	13.755	17.835	0.387
Baseline + AlignQualityScore	0.367	13.634	17.683	0.403
Baseline + POSAlignments	0.347	13.767	17.838	0.385
Baseline + BE_SentSim	0.348	13.756	17.828	0.387
Baseline + BE_WordSim	0.348	13.753	17.831	0.387
Baseline + BE_NounSim	0.346	13.780	17.857	0.385
Baseline + Ngrams	0.366	13.663	17.705	0.402
All Features	**0.377**	**13.603**	**17.642**	**0.410**

Table 4: Evaluation of the introduced features using WMT16 Sentence-Level QE Test set

6 Experiments

In our submission, we use the Python wrapper for BiSkip provided in the MultiVec tool[5] (Bérard et al., 2016). To train the model, we use the ITcorpus with the default configuration of the tool. The model was trained using a learning rate α set to 0.05 and $sample$ (a threshold on words' frequency) set to 0.001.

As a prediction model, we use the Linear Regression model to predict the post-editing effort need for each translation. In our experiments, we tried different combinations of the introduced features. Best results are obtained by training the model using all the features.

Tables 3 and 4 list the results of examined feature combinations on the development and test parts of QEcorpus, respectively. (The golden truth of the test part was made available only after the outputs submission deadline.) The models are evaluated in terms of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Pearson's correlation (Pearson's r) for post-editing effort prediction, and Spearman's rank correlation coefficient (Spearman's ρ) for the ranking task.

Results show that adding the alignment quality score to the set of baseline features gives the best performance compared to the other introduced features on the test set.

When added alone, features based on POS tags or bilingual embeddings do not help and sometimes even slightly degrade the performance, but apparently, they are useful in the combination.

Our submission to the task corresponds to the line "All Features" in Table 4.

Additionally, we experimented with replacing the parallel ITcorpus with only the comparable (but larger) ComparableNews when extracting bilingual embeddings. As documented in Table 5, the size of the monolingual data is apparently more important for the quality of the alignments. MultiVec, given two corpora, extracts the word alignments automatically, and obviously, it is going to fail most of the time when given a non-parallel corpus. Nevertheless, the few random alignments are probably sufficient to blend the source and target subspaces of the vector representation of words, because the setup with all BE features trained on ComparableNews instead of ITcorpus works better.

BE source	Pearson's r	MAE	RMSE
ITcorpus	0.377	13.603	17.642
ComparableNews	0.386	13.552	17.552

Table 5: Results of All Features with bilingual embeddings trained on ITcorpus or ComparableNews

[5] https://github.com/eske/multivec

6.1 Official Results

The official results of the WMT16 Sentence-Level QE task use Pearson's correlation as the primary evaluation metric for Scoring sub-task and Spearman's rank correlation as the primary evaluation metric for Ranking sub-task.

According to the official evaluation, our model is ranked 7^{th} (out of 14) and 6^{th} (out of 11) in the scoring and ranking sub-tasks respectively. As illustrated in Tables 6 and 7, our model outperforms the baseline system as well as the Referential Translation Machine model (RTM), the best performing system in WMT15 (Bicici et al., 2015), in both scoring and ranking sub-tasks on WMT16 IT-domain datasets.

7 Conclusion

In this paper, we described our submission to the WMT16 Quality Estimation Shared Task for English-German sentence-level post editing effort prediction and ranking. We introduced a new set of system independent features using bilingual distributed representations, word alignments and also frequent n-grams appearing in manually post-edited texts. Combined with baseline features, our features show an improvement in the performance of post-editing effort prediction in QE task.

An interesting observation is that the bilingual embeddings perform better when trained on a larger but only comparable corpus than on an in-domain parallel corpus. The bilingual embeddings are not trained specifically for the QE prediction and their contribution is thus arguably limited.

In the future, we plan to investigate more variants to the core learning model as well as training the embeddings for the specific task.

8 Acknowledgement

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 645452 (QT21).

References

Alexandre Bérard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. 2016. MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP. In *The 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*.

Ergun Bicici, 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*. Association for Computational Linguistics, Lisbon, Portugal, pages 304–308.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence Estimation for Machine Translation. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, page 315.

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 *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, pages 1–44.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 12–58.

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*. Association for Computational Linguistics, Lisbon, Portugal, pages 1–46.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics* 19(2):263–311.

Chris Callison-Burch, Philipp Koehn, Christof

System	Pearson's r	MAE	RMSE
YSDA/SNTX+BLEU+SVM	0.525	12.30	16.41
POSTECH/SENT-RNN-QV2	0.460	13.58	18.60
SHEF/SVM-NN-both-emb-QuEst	0.451	12.88	17.03
POSTECH/SENT-RNN-QV3	0.447	13.52	18.38
SHEF/SVM-NN-both-emb	0.430	12.97	17.33
UGENT/SVM2	0.412	19.57	24.11
Our Model	**0.377**	**13.60**	**17.64**
RTM/RTM-FS-SVR	0.376	13.46	17.81
UU/UU-SVM	0.370	13.43	18.15
UGENT/SVM1	0.363	20.01	24.63
RTM/RTM-SVR	0.358	13.59	18.06
BASELINE	0.351	13.53	18.39
SHEF/SimpleNets-SRC	0.320	13.92	18.23
SHEF/SimpleNets-TGT	0.283	14.35	18.22

Table 6: Official results for WMT16 Sentence-Level QE Scoring sub-task

System	Spearman's rho	DeltaAvg
POSTECH/SENT-RNN-QV2	0.483	7.663
SHEF/SVM-NN-both-emb-QuEst	0.474	8.129
POSTECH/SENT-RNN-QV3	0.466	7.527
SHEF/SVM-NN-both-emb	0.452	7.886
UGENT/SVM2	0.418	7.615
Our Model	**0.410**	**7.114**
UU/UU-SVM	0.405	6.519
RTM/RTM-FS-SVR	0.400	6.655
BASELINE	0.390	6.298
RTM/RTM-SVR	0.384	6.379
UGENT/SVM1	0.375	7.008

Table 7: Official results for WMT16 Sentence-Level QE Ranking sub-task

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*. Association for Computational Linguistics, Montréal, Canada, pages 10–51.

José Guilherme Camargo de Souza, Christian Buck, Marco Turchi, and Matteo Negri. 2013. FBK-UEdin Participation to the WMT13 Quality Estimation Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, pages 352–358.

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An Autoencoder Approach to Learning Bilingual Word Representations. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 1853–1861.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2014. BiLBOWA: Fast Bilingual Distributed Representations without Word Alignments. *arXiv preprint arXiv:1410.2455* .

Christian Hänig, Robert Remus, and Xose de la Puente. 2015. ExB Themis: Extensive Feature Extraction from Word Alignments for Semantic Textual Similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, pages 264–268.

Karl Moritz Hermann and Phil Blunsom. 2014.

Multilingual Models for Compositional Distributed Semantics. *CoRR* abs/1404.4641.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word Representations with Monolingual Quality in Mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. pages 151–159.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 3111–3119.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1):19–51.

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

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

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*. Association for Computational Linguistics, Montréal, Canada, pages 145–151.

Lucia Specia, G Paetzold, and Carolina Scarton. 2015. Multi-level Translation Quality Prediction with QuEst++. In *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*. pages 115–120.

Lucia Specia, Kashif Shah, José GC De Souza, and Trevor Cohn. 2013. QuEst - A Translation Quality Estimation Framework. In *ACL (Conference System Demonstrations)*. Citeseer, pages 79–84.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *13th Annual Conference of the European Association for Machine Translation*. Barcelona, Spain, EAMT, pages 28–37.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. DLS@CU: Sentence Similarity from Word Alignment and Semantic Vector Composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, pages 148–153.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pages 173–180.

Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *EMNLP*. pages 1393–1398.