Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 256–268

Semantic Textual Similarity in Quality Estimation

Hanna BÉCHARA[1], Carla PARRA ESCARTÍN[2], Constantin ORĂSAN[1],
Lucia SPECIA[3]

[1] University of Wolverhampton, Wolverhampton, UK
[2] Hermes Traducciones, Madrid, Spain
[3] University of Sheffield, Sheffield, UK

Hanna.Bechara@wlv.ac.uk, carla.parra@hermestrans.com,
C.Orasan@wlv.ac.uk, l.specia@sheffield.ac.uk

Abstract. Quality Estimation (QE) predicts the quality of machine translation output without the need for a reference translation. This quality can be defined differently based on the task at hand. In an attempt to focus further on the adequacy and informativeness of translations, we integrate features of semantic similarity into QuEst, a framework for QE feature extraction. By using methods previously employed in Semantic Textual Similarity (STS) tasks, we use semantically similar sentences and their quality scores as features to estimate the quality of machine translated sentences. Preliminary experiments show that finding semantically similar sentences for some datasets is difficult and time-consuming. Therefore, we opt to start from the assumption that we already have access to semantically similar sentences. Our results show that this method can improve the prediction of machine translation quality for semantically similar sentences.

Keywords: Quality Estimation, Semantic Textual Similarity, Machine Translation

1 Introduction

Machine Translation Quality Estimation (MTQE) has been gaining increasing interest in Machine Translation (MT) output assessment, as it can be used to measure different aspects of correctness. Furthermore, Quality Estimation (QE) tools forego the need for a reference translation and instead predict the quality of the output based on the source.

In this paper, we address the use of semantic correctness in QE by integrating STS measures into the process, without relying on a reference translation. We propose a set of features that compares MT output to a semantically similar sentence, that has already been assessed, using monolingual STS tools to measure the semantic proximity of the sentence in relation to the second sentence.

The rest of this paper is organised as follows: Section 2 features the state of the art in QE and the context for our research. Section 3 introduces our approach to integrating semantic information into QE. Section 4 details our experimental set-up, including the tools we use for our experiments. Section 5 explains our experiments, details our new STS features and summarises the results we observe when adding these features to QuEst. Finally, Section 6 presents our concluding remarks and plans for future work.

2 Previous Work

Early work in QE built on the concept of confidence estimation used in speech recognition (Gandrabur and Foster, 2003, Blatz et al., 2004). These systems usually relied on system-dependent features, and focused on measuring how confident a given system is rather than how correct the translation is.

Later experiments in QE used only system-independent features based on the source sentence and target translation (Specia et al., 2009b). They trained a Support Vector Machine (SVM) regression model based on 74 shallow features, and reported significant gains in accuracy over MT evaluation metrics. At first, these approaches to QE focused mainly on shallow features based on the source and target sentences. Such features include n-gram counts, the average length of tokens, punctuation statistics and sentence length among other features. Later systems incorporate linguistic features such as part of speech tags, syntactic information and word alignment information (Specia et al., 2010).

In the context of QE, the term “quality” itself is flexible and can change to reflect specific applications, from quality assurance, gisting and estimating post-editing (PE) effort to ranking translations. Specia et al. (2009a) define quality in terms of PE efficiency, using QE to filter out sentences that would require too much time to post-edit. Similarly, He et al. (2010) use QE techniques to predict human PE effort and recommend MT outputs to Translation Memory (TM) users based on estimated PE effort. In contrast, Specia et al., 2010 use QE to rank translations from different systems and highlight inadequate segments for post-editing.

Since 2012, QE has been the focus of a shared task at the annual Workshop for Statistical Machine Translation (WMT) (Callison-Burch et al., 2012). This task has provided a common ground for the comparison and evaluation of different QE systems and data at the word, sentence and document level (Bojar et al., 2015).

There have been a few attempts to integrate semantic similarity into the MT evaluation (Lo and Wu, 2011, Castillo and Estrella, 2012). The results reported are generally positive, showing that semantic information is not only useful, but often necessary, in order to assess the quality of machine translation output.

Specia et al. (2011) bring semantic information into the realm of QE in order to address the problem of meaning preservation. The authors focus on what they term “adequacy indicators” and human annotations for adequacy. The results they report show improvement with respect to a majority class baseline. Rubino et al. (2013) also address MT adequacy using topic models for QE. By including topic model features that focus on content words in sentences, their system outperforms state-of-the-art approaches specifically with datasets annotated for adequacy. Biçici (2013) introduce the use of referential translation machines (RTM) for QE. RTM is a computational model for judging monolingual and bilingual similarity that achieves state-of-the-art results. The authors report top performance in both sentence level and word-level tasks of WMT

2013. Camargo de Souza et al. (2014) propose a set of features that explore word alignment information in order to address semantic relations between sentences. Their results show that POS indicator features improve over the baseline at the shared task for QE at the workshop for machine translation. Kaljahi et al. (2014) employ syntactic and semantic information in quality estimation and are able to improve over the baseline when combining these features with the surface features of the baseline. Our work builds on previous work, focusing on the necessity of semantic information for MT adequacy. As far as we are aware, our work is the first to explore quality scores from semantically similar sentences as surrogate to the quality of the current sentence.

3 Our Approach

In this paper, we propose integrating semantic similarity into the quality estimation task. As STS relies on monolingual data, we employ the use of a second sentence that bears some semantic resemblance to the sentence we wish to evaluate.

Our approach is illustrated in Figure 1, where sentences A and B are two semantically similar sentences with a similarity score R. Our task is to assess the quality of sentence A with the help of sentence B which has already undergone machine translation evaluation, either through post-editing or by human evaluation (e.g. assessed on a scale from 1–5). As both sentences, A and B are semantically similar, our hypothesis is that their translations are also semantically similar and thus we can use the reference of sentence B to estimate the quality of sentence A.

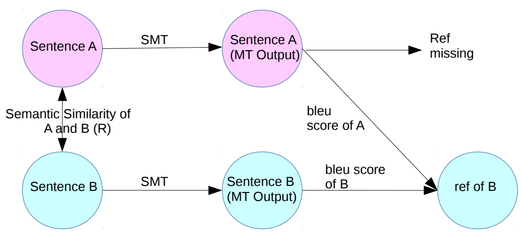

Fig. 1. Predicting the Quality of MT Output using a Semantically Similar Sentence B

For each sentence A, for which we wish to estimate MT quality, we retrieve a semantically similar sentence B which has been machine translated and has a reference translation or a quality assessment value. We then extract the following three scores (that we use as STS features):

Semantic Textual Similarity (STS) score: R represents the STS between the source sentence pairs (sentence A and sentence B). This is a continuous score ranging from 0

to 5. We calculate this score using the MiniExperts system designed for SemEval2015 (cf. Section 4.2) in all but one of our experiments, where we already have human annotations about STS. This experiment, the oracle experiment represents scores we could achieve if our STS was perfect.

Quality Score for Sentence B: We calculate the quality of the MT output of Sentence B. This is either a S-BLEU score based on a reference translation, or a manual score provided by a human evaluator.

S-BLEU Score for Sentence A: We have no human evaluation or reference translation for Sentence A, but we can calculate a quality score using Sentence B as a reference. We use sentence-level BLEU (S-BLEU) (Lin and Och, 2004). S-BLEU is designed to work at the sentence level and will still positively score segments that do not have a high order n-gram match.

4 Experimental Setting

In this section, we start with a brief introduction to the QuEst framework, followed by a description of the settings for the experiments described in this paper.

4.1 The QuEst Framework

QuEst (Specia et al., 2013) is an open source framework for MTQE.[4] In addition to a feature extraction framework, QuEst also provides the machine learning algorithms necessary to build the prediction models. QuEst gives access to a large variety of features, each relevant to different tasks and definitions of quality.

As QuEst is a state-of-the-art tool for MTQE and is used as a baseline in recent QE tasks, such as previous workshops for machine translation (Callison-Burch et al., 2012, Bojar et al., 2013, Bojar et al., 2014 and Bojar et al., 2015), we use its 17 features as a baseline to allow for comparison of our work to a state-of-the-art system.

The baseline features are system independent and include shallow surface features such as the number of punctuation marks, the average length of words and the number of words. Furthermore, these features include n-gram frequencies and language model probabilities. A full list of the baseline features can be found in Table 1.

4.2 MiniExpert's STS Tool

In our experiments, we use the MiniExpert's submission to Semeval2015's Task 2a (Béchara et al., 2015). The source code is easy to use and available on GitHub.[5] The system uses a SVM regression model to predict the STS scores between two English

[4] https://github.com/lspecia/quest
[5] https://github.com/rohitguptacs/wlvsimilarity

Table 1. Full List of QuEst's Baseline Features

ID	Description
1	number of tokens in the source sentence
2	number of tokens in the target sentence
3	average source token length
4	LM probability of source sentence
5	LM probability of the target sentence
6	average number of occurrences of the target word within the target sentence
7	average number of translations per source word in the sentence (as given by IBM 1 table thresholded so that prob(t|s) > 0.2)
8	average number of translations per source word in the sentence weighted by the inverse frequency of each word in the source corpus
9	percentage of unigrams in quartile 1 of frequency (lower frequency words) in a corpus of the source language
10	percentage of unigrams in quartile 4 of frequency (higher frequency words) in a corpus of the source language
11	percentage of bigrams in quartile 1 of frequency of source words in a corpus of the source language
12	percentage of bigrams in quartile 4 of frequency of source words in a corpus of the source language
13	percentage of trigrams in quartile 1 of frequency of source words in a corpus of the source language
14	percentage of trigrams in quartile 4 of frequency of source words in a corpus of the source language
15	percentage of unigrams in the source sentence seen in a corpus
16	number of punctuation marks in source sentence
17	number of punctuation marks in target sentence

sentences. The authors train their system on a variety of linguistically motivated features inspired by deep semantics with distributional Similarity Measures, Conceptual Similarity Measures, Semantic Similarity Measures and Corpus Pattern Analysis.

The system performs well and obtained a mean 0.7216 Pearson correlation in the shared task, ranking 33 out of 74 systems.

We train the STS tool on the SICK dataset Marelli et al., 2014, a dataset specifically designed for semantic similarity and used in previous SemEval tasks, augmented with training data from previous SemEval tasks (SemEval2014 and SemEval2015).

4.3 Statistical Machine Translation System

All of our experiments require MT output to run MTQE tasks. To that end, we use the state-of-the-art phrase based Statistical Machine Translation (SMT) system Moses (Koehn et al., 2007). We build 5-gram language models with Kneser-Ney smoothing trained with SRILM, (Stolcke, 2002), and run the GIZA++ implementation of IBM word alignment model 4 (Och and Ney, 2003), with refinement and phrase-extraction heuristics as described in Koehn et al. (2003). We use Minimum Error Rate Training (MERT) (Och, 2003) for tuning.

In order to keep our experiments consistent, we use the same SMT system for all datasets. We focus on English into French translations and we use the Europarl corpus (Koehn, 2005) for training. We train on 500,000 unique English–French sentences and then tune our system (using MERT) on 1,000 different unique sentences also from the Europarl corpus. We also train a French–English system to retrieve the backtranslations used in some of our experiments.

5 Experiments

As mentioned earlier, all our datasets focus on MTQE for English→French MT output. In all our experiments we have a set of machine translated sentences A for which we need a QE and a set of sentences B, semantically similar to the set of sentences A and for which we have some type of evaluation score available.

In early experiments, we attempted to use freely available datasets used in previous workshops on machine translation (WMT2012 and WMT2013) for the translation task and within the news domain (Bojar et al., 2013). The WMT datasets have two main advantages: first, they allow us to compare our system with previous systems for QE and render our experiments replicable. Second, they have manual evaluations that are available with the machine translations. Each sentence in the WMT dataset comes with a score between 1 and 5, provided by human annotators. However, this method proved to be too time-consuming, as it often required scoring thousands of sentences before finding two that were similar.

The first obstacle we faced in testing our approach with these datasets was the collection of similar sentences against which to compare and evaluate. We automatically searched large parallel corpora for sentences that yielded high similarity scores. These corpora included the Europarl corpus (Koehn, 2005), the Acquis Communautaire (Steinberger et al., 2006) and previous WMT data (from 2012 and 2013).

Furthermore, the STS system we use (see Section 4.2) returned many false-positives. Some sentences which appeared similar to the STS system were actually too different to be usable. This led to noisy data and unusable results. The scarcity of semantically similar sentences and the computational cost of finding these sentences, lead us to look into alternate datasets, preferably those with semantic similarity built into the corpus: the DGT-TM and the SICK dataset.

All our experiments have the same set-up. In all cases, we used 500 randomly selected sentences for testing, and the remaining sentences in the respective data-set for training QuEst. We automatically search large parallel corpora for sentences that yield high similarity scores using the STS system described in section 4.2.

We attempt to predict the quality scores of the individual sentences, using the STS features described above, added to QuEst's 17 baseline features. We compare our results to both the QuEst baseline (cf. Section 4.1). and the majority class baseline[6]. We also test our STS-related features separately, without the baseline features, and compare them to the system with the combined system (STS+baseline).

We use the Mean Absolute Error (MAE) to evaluate the prediction rate of our systems. MAE measures the average magnitude of the errors on the test set, without considering their direction. Therefore, it is ideal for measuring the accuracy for continuous variables. MAE is calculated as per Equation 1.

$$MAE = \frac{1}{n} \sum |x_i - y| \tag{1}$$

[6] The Mean Absolute Error calculated using the mean rating in the training set as a projected score for every sentence in the test set.

where n is the number of instances in the test set, x_i is the score predicted by the system, and y is the observed score. In our experiments, we use S-BLEU scores as the observed score.

5.1 DGT-TM

We use the 2014 release of the Directorate General for Translation – Translation Memory (DGT-TM) to test our system. The DGT-TM is a corpus of aligned sentences in 22 different languages created from the European Union's legislative documents (Steinberger et al., 2006). We randomly extract 500 unique sentences (B), then search the rest of the TM for the 5 most semantically similar sentences (A) for each of these 500 sentences (STS score > 3). This results in 2,500 sentences A (500x5) and their semantically similar sentence pairs B. We make sure to avoid any overlap in sentence A so that while semantically similar sentence B might recur, sentence A will remain unique. we assign an STS score to the resulting dataset using the system described in Section 4.2. We then translate these sentence pairs using the translation model described in Section 4.3 and use S-BLEU to assign evaluation scores for the MT outputs of Sentence A and B.

Of these 2,500 sentence pairs and their MT outputs, we use 2,000 sentence pairs to train an SVM regression model on Quest's baseline features using using sentence A and its MT output as the source and target sentence. We further use sentence B's S-BLEU score and its STS score with sentence A. We use the remaining 500 sentences to test our system. Table 2 shows a sample sentence (Sentence B) from the DGT-TM along its semantically similar retrieved match (Sentence A) and the machine translation output for each sentence. The MiniExpert's STS system gave the original English sentence pair a STS score of 4.46, indicating that only minor details differ.

Table 2. DGT-TM Sample Sentence

Sentence A	
Source	In order to ensure that the measures provided for in this Regulation are effective , it should enter into force *on the day of its publication*
MT	afin de garantir que les mesures prévues dans ce règlement sont efficaces , il devrait entrer en vigueur *sur le jour de sa publication ,*
Sentence B	
Source	In order to ensure that the measures provided for in this Regulation are effective , this Regulation should enter into force *immediately* ,
MT	afin de garantir que les mesures prévues dans ce règlement sont efficaces , ce règlement doit entrer en vigueur *immédiatement ,*
STS	4.46

Results: Our results are summarised in Table 3, which shows that the MAE for the combined features (QuEst + STS features) is considerably lower than that of QuEst on its own. This means that the additional use of STS features can improve QuEst's predictive power. Even the 3 STS features on their own outperformed QuEst's baseline features. These results show that our method can prove useful in a context where semantically similar sentences are accessible.

Table 3. Predicting the S-BLEU scores for DGT-TM - Mean Absolute Error

	MAE
QuEst Baseline (17 Features)	0.120
STS (3 Features)	0.108
Combined (20 Features)	0.090

5.2 SICK Dataset

In order to further test the suitability of our approach for semantically similar sentences, we use the SICK dataset for further experiments. SICK (Sentences Involving Compositional Knowledge) is a dataset specifically designed for compositional distributional semantics. It includes a large number of English sentence pairs that are rich in lexical, syntactic and semantic phenomena. The SICK dataset is generated from existing datasets based on images and video descriptions, and each sentence pair is annotated for relatedness (similarity) and entailment by means of crowd-sourcing techniques (Marelli et al., 2014). This means that we did not need to use the STS tool to annotate the sentences. The similarity score is a score between 1 and 5, further described in Table 4. As these scores are obtained by averaging several separate annotations by distinct evaluators, they are continuous, rather than discrete. As SICK already provides us with sentence pairs of variable similarity, it cuts out the need to search extensively for similar sentences. Furthermore, the crowd-sourced similarity scores act as a gold standard that eliminates the uncertainty introduced by the automatic STS tool. This dataset lacks a reliable reference translation to compare against, however.

Table 4. STS scale used by SemEval

0	The two sentences are on different topics
1	The two sentences are not equivalent, but are on the same topic
2	The two sentences are not equivalent, but share some details
3	The two sentences are roughly equivalent, but some important information differs/is missing
4	The two sentences are mostly equivalent, but some unimportant detail differs/missing
5	The two sentences are completely equivalent, as they mean the same thing

We extract 5,000 sentence pairs to use in our experiments and translate them into French using the MT system described in Section 4.3. The resulting dataset consists of 5,000 semantically similar sentence pairs and their French machine translations. Of this set, 4,500 are used to train an SVM regression model in the same manner as described in Section 5.1. The remaining 500 sentences are used for testing.

As the SICK dataset is monolingual and therefore lacking in a reference translation, we opted to use a back-translation (into English) as a reference instead of a French translation for these results. A back-translation is a translation of a translated text back into the original language. Back-translations are usually used to compare translations with the original text for quality and accuracy, and can help to evaluate equivalence of meaning between the source and target texts. In machine translation contexts, they can be used to create a pseudo-source that can be compared against the original source. He et al. (2010) used this back-translation as a feature in QE with some success. They compared the back-translation to the original source using fuzzy match scoring and used the result to estimate the quality of the translation. The intuition here is that the closer the back translation is to the original source, the better the translation is in the first place.

Following this idea, we use the S-BLEU scores of the back-translations as stand-ins for the MT quality scores. We use the MT system described in Section 3.3 for the back-translations.

Table 5 shows a sample sentence from the resulting dataset, including the original English sentence pairs and each sentence's MT output. The crowd-sourced STS score for this sentence pair is 4, indicating that only minor details differ.

Table 5. SICK Sample Sentence

Sentence A	
Source	Several children are *lying* down and are raising their knees
MT	Plusieurs enfants sont couchés et élèvent leurs genoux
Sentence B	
Source	Several children are *sitting* down and have their knees raised
MT	Plusieurs enfants sont assis et ont soulevé leurs genoux
STS	4

Results: Results on the SICK datasets are summarised in Table 6. The lowest error rate (MAE) is observed for the system that combined our STS-based features with QuEst's baseline features (Combined (20 Features)) just as in the DGT-TM experiments. We observe that even the STS features on their own outperformed QuEst in this environment.

Table 6. Predicting the S-BLEU scores for SICK - Mean Absolute Error

	MAE
QuEst Baseline (17 Features)	0.200
STS (3 Features)	0.189
Combined (20 Features)	0.177

The cherry-picked examples in Table 7 are from the SICK dataset, and show that a high STS score between the source sentences can contribute to a high prediction accuracy. In both examples, the predicted score for Sentence A is close to the actual observed score.

Table 7. SICK Sample Prediction

	Sentence A	Sentence B
Source MT	Dirt bikers are riding on a trail Dirt Bikers roulent sur une piste	Two people are riding motorbikes Deux personnes font du vélo motos
S-BLEU:	0.55 (Predicted) 0.6 (Actual)	0.84
STS	3.6	
Source MT S-BLEU:	A man is leaning against a pole and is surrounded Un homme est appuyée contre un poteau et est entouré par des gens 0.91 (Predicted) 1 (Actual)	A man is leaning against a pole and is surrounded by people Un homme est appuyée contre un poteau et est entouré 0.91
STS	4.2	

Furthermore, when we filtered the test set for the SICK experiments for sentences with high similarity (4+), we observed an even higher drop in MAE, as demonstrated in Table 8. This suggests that our experiments perform especially well if we select for sentences with high similarity.

Table 8. Predicting the S-BLEU scores for SICK sentences with high similarity - Mean Absolute Error

	MAE
QuEst Baseline (17 Features)	0.20
Combined (20 Features)	0.15

6 Conclusion and Future Work

In this paper we presented 3 semantically motivated features that augment QuEst's baseline features. We tested our approach on three different datasets and the results are encouraging, showing that these features can improve over the baseline when a sufficiently similar sentence against which to compare is available.

Several factors can be enhanced to further improve our system. To start with, the use of S-BLEU to evaluate our system is not ideal. Criticisms of BLEU and n-gram matching metrics in general are addressed by Callison-Burch et al. (2008), who show that BLEU fails to correlate to (and even contradicts) human judgement. More importantly, BLEU itself does not measure meaning preservation. Therefore, to evaluate our system more thoroughly, we would need to compare it to human judgements. In order to address the criticisms of both BLEU and the back-translations, we are currently collecting manual evaluations of the French SICK MT output sentences. Before a full manual evaluation is performed, we cannot conclusively state that our results on the SICK dataset are valid in a real world setting.

Another case worth addressing further is that where the retrieved matches are so similar to the original, that they could be acting as a pseudo-reference. While the examples show that this is not always the case, this phenomenon bears further investigation in future research.

The MiniExpert's tool which we use to determine the STS scores for the DGT–TM is trained on very different data (the SICK corpus and SemEval data), which may affect its accuracy. This may explain why it did not work as well as expected given its reported performance. However, the lack of readily available semantically annotated data to train on limits us in this regard. Furthermore, our features rely on the existence of semantically similar sentences against which we can compare our translations. These sentences are not always readily available and, as explained earlier in Section 5, searching large corpora for similar sentences can be computationally costly and time-consuming.

In spite of these short-comings, this approach can be quite useful in settings where we wish to predict the quality of sentences within a very specific domain. One potential such scenario, would be post-editing tasks in which professional translators are asked to post-edit MT output of specialized texts. As translators use Translation Memories (TMs) to ensure the quality of their work, such TMs could be used to obtain semantically similar sentences to the ones in the MTPE task and compute with our approach a QE score. The results we obtained in the case of SICK are encouraging in this respect and in future work we plan to investigate this further.

Acknowledgements

This work is 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

Béchara, H., Costa, H., Taslimipoor, S., Gupta, R., Orasan, C., Corpas Pastor, G., and Mitkov, R. (2015). MiniExperts: An SVM approach for Measuring Semantic Textual Similarity. In 9^{th} *Int. Workshop on Semantic Evaluation*, SemEval'15, pages 96–101, Denver, Colorado. ACL.

Biçici, Ergun (2013). Referential translation machines for quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 343–351, Sofia, Bulgaria.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, C., Sanchis, A., and Ueffing, N. (2004). Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (CoLing-2004)*, pages 315–321.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

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

Bojar, O. and Chatterjee, R. and Federmann, C. and Haddow, B. and Huck, M. and Hokamp, C. and Koehn, P. and Logacheva, V. and Monz, C. and Negri, M. and others 2015 Findings of the 2015 Workshop on Statistical Machine Translation

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further Meta-Evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT)*, pages 70–106.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L., editors (2012). *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Montréal, Canada.

Castillo, J. and Estrella, P. (2012). Semantic textual similarity for mt evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 52–58, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gandrabur, S. and Foster, G. (2003). Confidence estimation for translation prediction. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 95–102, Stroudsburg, PA, USA. Association for Computational Linguistics.

He, Y., Ma, Y., van Genabith, J., and Way, A. (2010). Bridging SMT and TM with Translation Recommendation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 622–630.

Kaljahi, R. and Foster, J. and Roturier, J. (2014). Syntax and Semantics in Quality Estimation of Machine Translation, In *Syntax, Semantics and Structure in Statistical Translation*, pages 67.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 177–180.

Koehn, P., Och, F., and Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 48–54.

Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lo, C.-k. and Wu, D. (2011). Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 220–229. Association for Computational Linguistics.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014b). A sick cure for the evaluation of compositional distributional semantic models. In *LREC'14*, Reykjavik, Iceland.

Och, F. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 160–167.

Och, F. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. In *Proceedings of the Association for Computer Linguistics (ACL)*, pages 29(1):19–51.

Rubino, Raphael and Souza, José Guilherme Camargo and Foster, Jennifer and Specia, Lucia. Topic Models for Translation Quality Estimation for Gisting Purposes. In Machine Translation Summit XIV, pages 295–302.

, Camargo de Souza, José Guilherme and González-Rubio, Jesús and Buck, Christian and Turchi, Marco and Negri, Matteo. FBK-UPV-UEdin participation in the WMT14 Quality Estimation shared-task. In Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 322–328.

Specia, L., Raj, D., and Turchi, M. (2010). Machine Translation Evaluation versus Quality Estimation. In *Machine Translation Volume 24, Issue 1*, pages 39–50.

Lucia Specia and Najeh Hajlaoui and Catalina Hallett and Wilker Aziz. Predicting Machine Translation Adequacy. In Machine Translation Summit XIII, pages 513–520.

Specia, L., Shah, K., De Souza, J. G. C., and Cohn, T. (2013). QuEst - A translation quality estimation framework. In *Proceedings of the Association for Computational Linguistics (ACL), Demonstrations*.

Specia, L., Turchi, M., Cancedda, N., Dymetman, M., and Cristianini, N. (2009a). Estimating the Sentence-Level Quality of Machine Translation Systems. In *13th Annual Meeting of the European Association for Machine Translation (EAMT-2009)*, pages 28–35.

Specia, L., Turchi, M., Wang, Z., Shawe-Taylor, J., and Saunders, C. (2009b). Improving the confidence of machine translation quality estimates.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., and Tufis, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC–2006*, pages 2142–2147.

Stolcke, A. (2002). SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904.

Received May3, 2016 , accepted May 13, 2016