Document-level translation quality estimation: exploring discourse and pseudo-references

Carolina Scarton and **Lucia Specia**
Department of Computer Science
University of Sheffield
S1 4DP, UK
{c.scarton,l.specia}@sheffield.ac.uk

Abstract

Predicting the quality of machine translations is a challenging topic. Quality estimation (QE) of translations is based on features of the source and target texts (without the need for human references), and on supervised machine learning methods to build prediction models. Engineering well-performing features is therefore crucial in QE modelling. Several features have been used so far, but they tend to explore very short contexts within sentence boundaries. In addition, most work has targeted sentence-level quality prediction. In this paper, we focus on document-level QE using novel discursive features, as well as exploiting pseudo-reference translations. Experiments with features extracted from pseudo-references led to the best results, but the discursive features also proved promising.

1 Introduction

The purpose of machine translation (MT) **quality estimation (QE)** is to provide a quality prediction for new, unseen machine translated texts, without relying on reference translations (Blatz et al., 2004; Specia et al., 2009; Bojar et al., 2013). This task is usually addressed with machine learning models trained on datasets composed of source texts, their machine translations, and a quality label assigned by humans or by an automatic metric (e.g.: BLEU (Papineni et al., 2002)). A common use of quality predictions is the estimation of post-editing effort in order to decide whether to translate a text from scratch or post-edit its machine translation. Another use is the ranking of translations in order to select the best text from multiple MT systems.

Feature engineering is an important component in QE. Although several feature sets have already been explored, most approaches focus on sentence-level quality prediction, with sentence-level features. These disregard document structure or wider contexts beyond sentence boundaries. To the best of our knowledge, only Rubino et al. (2013) considered discourse-related information by studying topic model features for sentence-level prediction. Soricut and Echihabi (2010) explored document-level quality prediction, but they did not use explicit discourse information, e.g. information to capture text cohesion or coherence.

In this paper we focus on **document-level features** and **document-level prediction**. We believe that judgements on translation quality depend on units longer than just a given sentence, taking into account discourse phenomena for lexical choice, consistency, style and connectives, among others (Carpuat and Simard, 2012). This is particularly important in MT evaluation contexts, since most MT systems, and in particular statistical MT (SMT) systems, process sentences one by one, in isolation. Our hypothesis is that features that capture **discourse phenomena** can improve document-level prediction. We consider two families of features that have been successfully applied in reference-based MT evaluation (Wong and Kit, 2012) and readability assessment (Graesser et al., 2004). In terms of applications, document-level QE is very important in scenarios where the entire text needs to be used/published without post-edition.

Soricut and Echihabi (2010) and Soricut and Narsale (2012) explored a feature based on **pseudo-references** for document-level QE. Pseudo-references are translations produced by one or more external MT systems, which are different from the one producing the translations we want to predict the quality for. These are used as references against which the output of the MT system of interest can be compared using standard metrics such as BLEU. Soricut et al. (2012) and Shah et al. (2013) explored pseudo-references for sentence-level QE. In both cases, features based on pseudo-references led to significant improvements in prediction accuracy. Here we also use pseudo-references for document-level QE, with a number of string similarity metrics to produce document-level scores as features, which are arguably more reliable than sentence-level scores, particularly for metrics like BLEU.

In the remainder of this paper, Section 2 presents related work. Section 3 introduces the document-level QE features we propose. Section 4 describes the experimental setup of this work. Section 5 presents the results.

2 Related work

Work related to this research includes document-level MT evaluation metrics, QE features, and QE prediction, as well as work focusing on other linguistic features, and work using pseudo-references.

Wong and Kit (2012) use lexical cohesion metrics for MT evaluation at document-level. Lexical cohesion relates to word choices, captured in their work by reiteration and collocation. Words and stems were used for reiteration, and synonyms, near-synonyms and superordinates, for collocations. These metrics are integrated with traditional metrics like BLEU, TER (Snover et al., 2006) and METEOR (Banerjee and Lavie, 2005). The highest correlation against human assessments was found for the combination of METEOR and the discursive features.

Rubino et al. (2013) explore topic model features for QE at sentence-level. Latent Dirichlet Allocation is used to model the topics in two ways: a bilingual view, where the bilingual corpus is concatenated at sentence-level to build a single model with two languages; and a polylingual view, where one topic model is built for each language. While the topics models are generated with information from the entire corpus, the features are extracted at sentence-level. These are computed for both source and target languages using vector distance metrics between the words in these sentences and the topic distributions. Topic model features has been achieved promising results.

Soricut and Echihabi (2010) explore document-level QE prediction to rank documents translated by a given MT system. Features included BLEU scores based on pseudo-references from an off-the-shelf MT system, for both the target and the source languages. The use of pseudo-references has been shown to improve state-of-the-art results. Soricut and Narsale (2012) also consider document-level prediction for ranking, proposing the aggregation of sentence-level features for document-level prediction. The authors claim that a pseudo-references-based feature (based in BLEU) is one of the most powerful in the framework. For QE at sentence-level, Soricut et al. (2012) use BLEU based on pseudo-references combined with other features to build the best QE system of the WMT12 QE shared task.[1] Shah et al. (2013) conduct a feature analysis, at sentence-level, on a number of datasets and show that the BLEU-based pseudo-reference feature contributes the most to prediction performance.

In terms of other types of linguistic features for QE, Xiong et al. (2010) and Bach et al. (2011) propose features for word-level QE and show that these improve over the state-of-the-art results. At sentence-level, Avramidis et al. (2011), Hardmeier (2011) and Almaghout and Specia (2013) consider syntactic features, achieving better results compared to competitive feature sets. Pighin and Màrquez (2011) obtain improvements over strong baselines from exploiting semantic role labelling to score MT outputs at sentence-level. Felice and Specia (2012) introduce several linguistic features for QE at sentence-level. These did not show improvement over shallower features, but feature selection analysis showed that linguistic features were among the best performing ones.

3 Features for document-level QE

QE is traditionally done at sentence-level. This happens mainly because the majority of MT systems translate texts at this level. Evaluating sentences instead of documents can be useful for many scenarios, e.g., post-editing effort prediction.

[1] http://www.statmt.org/wmt12/

However, some linguistic phenomena can only be captured by considering the document as a whole. Moreover, for scenarios in which post-edition is not possible, e.g., gisting, quality predictions for the entire documents are more useful.

Several features have been proposed for QE at sentence-level. Many of them can be directly used at document-level (e.g., number of words in source/target sentences). However, other features that better explore the document as a whole or discourse-related phenomena can bring additional information. In this paper, discourse information is explored in two ways: lexical cohesion (Section 3.1) and LSA cohesion (Section 3.2). The intuition behind using cohesion features for QE is the following: on the source side, documents that have low cohesion are likely to result in bad quality translations. On the target side, documents with low cohesion are likely to have low overall quality.

From the feature set proposed in (Soricut and Echihabi, 2010) for document-level ranking of MT system outputs, text-based and language model-based features are also covered by the baseline features used in this paper. Pseudo-reference-based features are also addressed herein (Section 3.3). The example-based features cannot be easily reproduced since we do not have access to additional documents to use as development set (our parallel corpora are already small). The training data-based features were not considered because we use MT systems that do not have or make their training sets available.

3.1 Lexical cohesion features

Our first set of features is based on lexical cohesion metrics (hereafter, **LC**). Lexical cohesion is related to word choices in a text (Wong and Kit, 2012). Words can be repeated to make the relation among sentences more explicit to the reader. Another phenomenon of lexical cohesion is the use of synonyms, hypernyms, antonyms, etc. In this paper, we only consider word repetitions as features. These are features that can be easily extracted for languages other than English, for which a thesaurus with synonyms, hypernyms, etc., may not be available. Our LC features are as follows:

Average word repetition: for each content word, we count its frequency in all sentences of the document. Then, we sum the repetition counts and divide it by the total number of content words in the document. This is computed for the source and target documents, resulting in two features.

Average lemma repetition: the same as above, but the words are first lemmatised.

Average noun repetition: the same as above, but only nouns are considered as words.

3.2 LSA cohesion features

General textual quality is often connected to the notion of readability of a text. Readability can be measured in many ways, focusing on different aspects such as coherence, cohesion, how accessible a text is to a certain audience, etc. The Coh-Metrix project[2] (Graesser et al., 2004) has proposed a number of text readability metrics. Latent Semantic Analysis (**LSA**) (Landauer et al., 1998) is used in order to extract cohesion-related features. This is a statistical method based on Singular Vector Decomposition (SVD) and is often aimed at dimensionality reduction. In SVD, a given matrix X can be decomposed into the product of three other matrices:

$$X = WSP^T,$$

where W describes the original row entities as vectors of derived orthogonal factor values; S is a diagonal matrix containing scaling values and P (P^T is the transpose of P) is the same as W but for columns. When these three matrices are multiplied, the exact X matrix is recovered. The dimensionality reduction consists in reconstructing the X matrix by only using the highest values of the diagonal matrix S. For example, a dimensionality reduction of order two will consider only the two highest values of S.

The X matrix (rows x columns) can be built from words by sentences, words by documents, sentences by documents, etc. In the case of words by sentences (which we use in our experiments), each cell contains the frequency of a given word in a given sentence. LSA was originally designed to be used with large corpora of multiple documents. In our case, since we are interested in measuring cohesion within documents, we compute LSA for each individual document through a matrix of words by sentences within the document.

LSA was computed using a package for python,[3] which takes word stems and sentences to build the matrix. Usually, before applying SVD in

[2] http://cohmetrix.com/

[3] https://github.com/josephwilk/semanticpy

LSA, the X matrix is transformed wherefore each cell encapsulates information about a word's importance in a sentence or a word's importance in the text in general. Landauer et al. (1998) suggest the use of TF-IDF transformation for that. However, we disregarded the use of TF-IDF as this transformation would smooth out the values of high frequency words across sentences. In our case, the salience of words in sentences is important.

Our LSA features follow from Graesser et al. (2004)'s work on readability assessment:

LSA adjacent sentences: for each sentence in a document, we compute the Spearman rank correlation coefficient of its word vector with the word vectors of its immediate neighbours (sentences which appear immediately before and after the given sentence). For sentences with two neighbours (most cases), we average the correlation values. After that, we average the values for all sentences in order to have a single figure for the entire document.

LSA all sentences: for each sentence in a document, we calculate the Spearman rank correlation coefficient of the word vectors between this sentence and all the others. Again we average the values for all sentences in the document.

Higher correlation scores are expected to correspond to higher text cohesion, since the correlation among the sentences in a document is related to how close the words in the document are (Graesser et al., 2004). Different from lexical cohesion features, LSA features are able to find correlations among different words, which are not repetitions and may not be synonyms, but are instead related (as given by co-occurrence patterns).

3.3 Pseudo-references

Pseudo-references are translations produced by other MT systems than the system we want to predict the quality for. They are used as references to evaluate the output of the MT system of interest. They have also been used for other purposes, e.g., to fulfil the lack of human references available in reference-based MT evaluation (Albrecht and Hwa, 2008) and automatic summary evaluation (Louis and Nenkova, 2013). The application we are interested in, originally proposed in (Soricut and Echihabi, 2010), is to generate features for QE. In this scenario, reference-based evaluation metrics (such as BLEU) are computed between the MT system output and the pseudo-references and used to train quality prediction models.

Soricut and Echihabi (2010) discussed the importance of the pseudo-references being generated by MT system(s) which are as different as possible from the MT system of interest, and preferably of much better quality. This should ensure that string similarity features (like BLEU) indicate more than simple consensus between similar MT systems, which would produce the same (possibly bad quality) translations, e.g., Google Translate[4].

4 Experimental settings

Although QE is traditionally trained on datasets with human labels for quality (such as HTER – Human Translation Error Rate (Snover et al., 2006)), no large enough dataset with human-based quality labels assigned at document-level is available. Therefore, we resort to predicting automatic metrics as quality labels, as in (Soricut and Echihabi, 2010). This requires references (human) translations at training time, when the automatic metrics are computed, but not at test time, when the automatic metrics are predicted.

Corpora Two parallel corpora with reference translations are used in our experiments: FAPESP and WMT13. **FAPESP** contains $2,823$ English-Brazilian Portuguese (EN-BP) documents extracted from a scientific Brazilian news journal (FAPESP)[5] (Aziz and Specia, 2011). Each article covers one particular scientific news topic. The corpus was randomly divided into 60% ($1,694$ documents) for training a baseline **MOSES**[6] statistical MT system (Koehn et al., 2007) (with 20 documents as development set); and 40% ($1,128$ documents) for testing the SMT system, which generated translations for QE training (60%: 677 documents) and test (40%: 451 documents). In addition, two external MT systems were used to translate the test set: **SYSTRAN**[7] – a rule-based system – and Google Translate (**GOOGLE**), a statistical system.

WMT13 contains English-Spanish (**EN-ES**) and Spanish-English (**ES-EN**) translations from

[4] http://translate.google.com.br/
[5] http://revistapesquisa.fapesp.br
[6] http://www.statmt.org/moses/?n=moses.baseline
[7] http://www.systransoft.com/

the test set of the translation shared task of WMT13.[8] In total, 52 source documents were available for each language pair. In order to build the QE systems, the outputs of all MT systems submitted to the shared task were taken: 18 systems for EN-ES (528 documents for QE training, and 356 for QE test), and 17 systems for ES-EN (500 documents for QE training, and 332 documents for QE test). In both cases, the translations from one MT system are used as pseudo-references for translations from the other systems.

Quality labels The automatic metrics selected for quality labelling and prediction are BLEU and TER.[9] **BLEU** (BiLingual Evaluation Understudy) is a precision-oriented metric that compares n-grams (n=1-4 in our case) from reference documents against n-grams of the MT output, measuring how close the output of the system is to one or more references. **TER** (Translation Error Rate) (Snover et al., 2006) measures the minimum number of edits required to transform the MT output in the reference document. The Asiya Toolkit[10](Giménez and Màrquez, 2010) was used to calculate both metrics.

Baselines As baseline, we use 17 competitive features from the QuEst toolkit (Specia et al., 2013) (the so-called **baseline features** or **BL**.[11]) Since the baseline features are sentence-level, we aggregated them by computing the average for each feature across all sentences in a document. As a second baseline (**Mean**), we calculate the average BLEU or TER scores in the QE training set, and apply this value to all entries (documents) in the test set.

Pseudo-reference features BLEU and TER scores are computed between the output of the MT system of interest and alternative MT systems, at document-level, and used as features in QE models. For the FAPESP corpus, translations from Google Translate were selected as pseudo-references, since this system has shown the best average BLEU score in the QE training set. For the WMT13 corpus, translations from *uedin-wmt13-en-es*, for EN-ES, and *uedin-heafield-unconstrained* for ES-EN, were used as pseudo-references, since these systems achieved the best BLEU scores in the WMT13 translation shared task. Regarding the difference between the systems, for the FAPESP corpus, this difference is guaranteed since GOOGLE is considerably different from SYSTRAN, and is trained on a different (much larger) corpus than MOSES. For the WMT13 corpus, it is not possible to make this assumption, as many of the systems participating in the shared task are close variations of Moses.

Feature sets As feature sets, we combine LC and LSA features with BL (**BL+LC**, **BL+LSA** and **BL+LC+LSA**) to create the models with discursive information. The pseudo-reference features are combined with the baseline (**BL+Pseudo**) and with all other features (**BL+LC+LSA+Pseudo**).

Machine learning algorithm We use the Support Vector Machines (SVM) regression algorithm with a radial basis function kernel and hyperparameters optimised via grid search to train the QE models with all feature sets The scikit-learn module available in QuEst was used for that.

Evaluation metrics The QE models with different feature sets are evaluated using **MAE** (Mean Absolute Error): $MAE = \frac{\sum_{i=1}^{n} |H(s_i) - V(s_i)|}{N}$ where $H(s_i)$ is the predicted score, $V(s_i)$ is the true score and N is the number of data points in the test set. To verify the significance of the results, two-tailed pairwise t-test ($p<0.05$) was performed for different prediction outputs.

Method Two sets of experiments were conduct. First (Section 5.1), we consider the outputs of the FAPESP corpus of MOSES, SYSTRAN and GOOGLE separately, using as training and test sets the outputs of each system individually, with GOOGLE translations used as pseudo-references for the other two systems. The second set of experiments (Section 5.2) considers, for the FAPESP corpus, the combination of the outputs of MOSES and SYSTRAN (MOS+SYS), again with GOOGLE translations used as pseudo-references. For the WMT2013 corpora, we mixed translations from all except the best system, which were used as pseudo-references.

5 Experiments and results

5.1 MT system-specific models

The results for the prediction of BLEU and TER for MOSES, SYSTRAN and GOOGLE systems

[8] http://www.statmt.org/wmt13/

[9] METEOR was also used but the results were inconclusive

[10] http://asiya.lsi.upc.edu/

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

in the FAPESP corpus are shown in Table 1. The best results for MOSES and SYSTRAN were obtained with the inclusion of pseudo-references (BL+Pseudo and BL+LC+LSA+Pseudo), with both BLEU and TER. However, only the improvements for MOSES showed statistically significant difference: with both BLEU and TER, the best results were tied between BL+Pseudo and BL+LC+LSA+Pseudo, but there are still significant differences between their predictions. An interesting finding is that without considering pseudo-reference features for MOSES and SYSTRAN, the best results are achieved with LSA features. In fact, for SYSTRAN the results from using of only BL+LSA are not significantly different from the use of all features (including pseudo-references).

For GOOGLE, the best results (for BLEU and TER) were obtained by BL+LC [12]. However, BLEU predictions showed no significant difference among all feature sets and the best TER figure was not significantly different from BL+LC+LSA.

In order to understand whether the MAE scores obtained are "good enough", it is interesting to compare them against the error of the Mean baseline, but also to analyse the average of the true scores and the range of variation of these true scores in the test set (last two lines in Table 1). For the prediction of BLEU scores, the true scores range from 0 to 0.5 for MOSES and SYSTRAN, and from 0 to 0.8 for GOOGLE. This suggests that the impact of error differences in MOSES and SYSTRAN is higher. A wider range of scores and a relatively higher Mean MAE could indicate a relatively easier prediction task. This is directly connected to the variation in the quality of the translations in the datasets. This seems to be the case with BLEU prediction for GOOGLE translations: the improvements between the Mean baseline and the BL features is much higher than with the other MT systems. The variation in terms of TER is larger, making improvements over the Mean baseline possible with all feature sets.

Given the low MAE scores obtained by the Mean baseline, as well as with simple BL features, one could say that in general the task of predicting BLEU and TER is close to trivial, at least in the FAPESP corpus. This is again due to the low variation in the quality of texts translated by each system. This is to be expected, given the very nature of document-level prediction: major variations in the quality of specific translated segments get smoothed out throughout the document. In addition, the FAPESP corpus consists of texts from the same style and domain. On the other hand, the average quality (as measured by BLEU and TER metrics) of the different MT systems on the same corpus is very different, as shown in the penultimate line of Table 1. This motivates the experiment described next.

5.2 MT system-independent models

To analyse document-level QE in a more challenging scenario, we experiment with mixing different MT system outputs, for both FAPESP and WMT2013 corpora. Results are shown in Table 2.

The ranges of BLEU/TER scores are now wider, and the overall error scores (including for the Mean baseline) are higher in these settings, showing that this is indeed a harder task. Again, the best results are obtained with the use of pseudo-reference features. However, in this case statistically significant differences against other results were only observed with MOS+SYS BLEU prediction and ES-EN TER prediction. For EN-ES BLEU prediction, the best result (0.043 for BL+Pseudo) showed no significant difference against BL+LC+LSA+Pseudo (0.045). For ES-EN BLEU prediction, there is no significant difference among the results of BL+LSA, BL+LC+LSA and BL+Pseudo. For MOS+SYS TER prediction, BL+Pseudo and BL+LC+LSA+Pseudo showed no significant difference. EN-ES TER prediction was the only case were the BL results showed no significant difference against pseudo-reference features. It is worth mentioning that, as in the previous experiments, if we disregard the pseudo-reference features – which may not be available in many real-world scenarios – the LSA feature sets show the best results.

6 Conclusions

In this paper we focused document-level machine translation quality estimation. We presented an attempt to address the problem by considering discourse information in translation quality estimation in terms of novel features, relying on lexical cohesion aspects. LSA cohesion features showed very promising results.

Features based on pseudo-references were also

[12] Pseudo-reference features were not used for GOOGLE, since its outputs was used as pseudo-reference for the other systems.

	BLEU			TER		
	MOSES	SYSTRAN	GOOGLE	MOSES	SYSTRAN	GOOGLE
Mean	0.059	0.047	0.066	0.063	0.062	0.068
BL	0.046	0.047	0.056	0.054	0.059	0.061
BL+LC	0.044	0.043	**0.055**	0.053	0.059	**0.055**
BL+LSA	0.044	0.044	0.058	0.055	0.059	0.060
BL+LC+LSA	0.044	0.043	0.057	0.053	0.058	0.061
BL+Pseudo	**0.042***	0.038	-	**0.052***	**0.051**	-
BL+LC+LSA+Pseudo	**0.042***	**0.036**	-	**0.052***	**0.051**	-
Test-set average	0.365	0.275	0.456	0.427	0.506	0.372
Test-set range	[0.004,0.558]	[0,0.406]	[0.004, 0.79]	[0.245,1.056]	[0,1.071]	[0.12,1.084]

Table 1: MAE scores for document-level prediction of BLEU and TER for the FAPESP corpus. Bold-faced figures indicate the smallest MAE for a given test set; * indicates a statistically significant difference against all other results; underlined values indicate no significant difference against the best system.

	BLEU			TER		
	FAPESP	WMT2013		FAPESP	WMT2013	
	MOS+SYS	EN-ES	ES-EN	MOS+SYS	EN-ES	ES-EN
Mean	0.064	0.061	0.076	0.07	0.066	0.089
BL	0.045	0.056	0.065	0.063	0.059	0.069
BL+LC	0.044	0.058	0.065	0.063	0.066	0.07
BL+LSA	0.044	0.052	0.051	0.062	0.057	0.051
BL+LC+LSA	0.044	0.053	0.052	0.064	0.054	0.062
BL+Pseudo	0.043	**0.043**	**0.038**	**0.053**	**0.034**	**0.038***
BL+LC+LSA+Pseudo	**0.038***	0.045	0.043	0.054	**0.034**	0.04
Test-set average	0.32	0.266	0.261	0.466	0.524	0.55
Test-set range	[0,0.558]	[0.107,0.488]	[0.072,0.635]	[0,1.07]	[0.317,0.72]	[0.216,0.907]

Table 2: MAE scores for document-level prediction of BLEU and TER for the FAPESP corpus (mixing MOSES and SYSTRAN) and for the WMT2013 EN-ES and ES-EN corpora (mixing all but best system).

explored. Confirming the findings in (Soricut and Echihabi, 2010; Shah et al., 2013), these features were found responsible for the most significant improvements over strong baselines. However, in most settings, our proposed LSA cohesion features performed as well as pseudo-reference features.

Predicting automatic metrics at document-level proved a less challenging task than we expected. This was mostly due to the low variance in the quality of translations for the various documents in the corpus by a given MT system. This was confirmed by the low prediction error obtained by a simple baseline that assigns the mean quality score (BLEU or TER) of the training set to all instances of the test set. Outperforming this mean baseline proved particularly difficult for some MT systems when predicting BLEU. Putting MT systems of various quality levels together made the task more complex. As a consequence, our QE models yielded more significant improvements over the baseline.

In future work, we plan to model this problem as predicting post-editing effort scores, as it has been done in the state-of-the-art work for QE at sentence-level. This will require larger datasets with post-edited machine translations and document-level markup.

Acknowledgements: This work was supported by the EXPERT (EU Marie Curie ITN No. 317471) project.

References

Albrecht, Joshua S. and Rebecca Hwa. 2008. The Role of Pseudo References in MT Evaluation. In *Proceedings of WMT 2008*, pages 187–190, Columbus, OH.

Almaghout, Hala and Lucia Specia. 2013. A CCG-based Quality Estimation Metric for Statistical Machine Translation. In *Proceedings of the XIV MT Summit*, pages 223–230, Nice, France.

Avramidis, Eleftherios, Maja Popovic, David Vilar Torres, and Aljoscha Burchardt. 2011. Evaluate with confidence estimation: Machine ranking of translation outputs using grammatical features. In *Proceedings of WMT 2011*, pages 65–70, Edinburgh, UK.

Aziz, Wilker and Lucia Specia. 2011. Fully Automatic Compilation of a Portuguese-English Parallel Corpus for Statistical Machine Translation. In *Proceedings of STIL 2011*, Cuiabá, MT, Brazil.

Bach, Nguyen, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: A method for measuring machine translation confidence. In *Proceedings of ACL 2011*, pages 211–219, Portland, OR.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI.

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

Bojar, Ondřej, 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 WMT 2013*, pages 1–44, Sofia, Bulgaria.

Carpuat, Marine and Michel Simard. 2012. The trouble with SMT consistency. In *Proceedings of WMT 2012*, pages 442–449, Montréal, Canada.

Felice, Mariano and Lucia Specia. 2012. Linguistic features for quality estimation. In *Proceedings of WMT 2012*, pages 96–103, Montréal, Canada.

Giménez, Jesús and Lluís Màrquez. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94:77–86.

Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36(2):193–202.

Hardmeier, Christian. 2011. Improving machine translation quality prediction with syntatic tree kernels. In *Proceedings of EAMT 2011*, pages 233–240, Leuven, Belgium.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Bertoldi Nicola Federico, Marcello, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. MOSES: Open source Toolkit for Statistical Machine Translation. In *Proceedings of ACL 2007, demonstration session*, Prague, Czech Republic.

Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2-3):259–284.

Louis, Annie and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics*, 39(2):267–300.

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

Pighin, D and L Màrquez. 2011. Automatic projection of semantic structures: an application to pairwise translation ranking. In *Proceedings of SSST-5*, pages 1–9, Portland, OR.

Rubino, Raphael, José G. C. de Souza, Jennifer Foster, and Lucia Specia. 2013. Topic Models for Translation Quality Estimation for Gisting Purposes. In *Proceedings of the XIV MT Summit*, pages 295–302, Nice, France.

Shah, Kashif, Trevor Cohn, and Lucia Specia. 2013. An Investigation on the Effectiveness of Features for Translation Quality Estimation. In *Proceedings of the XIV MT Summit*, pages 167–174, Nice, France.

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

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

Soricut, Radu 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.

Soricut, Radu, Nguyen Bach, and Ziyuan Wang. 2012. The SDL Language Weaver Systems in the WMT12 Quality Estimation Shared Task. In *Proceedings of WMT 2012*, pages 145–151, Montréal, Canada.

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

Specia, Lucia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - A translation quality estimation framework. In *Proceedings of WMT 2013: System Demonstrations*, ACL-2013, pages 79–84, Sofia, Bulgaria.

Wong, Billy T. M. and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of EMNLP-CONLL 2012*, pages 1060–1068, Jeju Island, Korea.

Xiong, Deyi, Min Zhang, and Haizhou Li. 2010. Error detection of statistical machine translation using linguistic features. In *Proceedings of ACL 2010*, pages 604–611, Uppsala, Sweden.