Exploiting Objective Annotations for Measuring Translation Post-editing Effort

Lucia Specia
Research Group in Computational Linguistics,
University of Wolverhampton,
Wolverhampton, UK
l.specia@wlv.ac.uk

Abstract

With the noticeable improvement in the overall quality of Machine Translation (MT) systems in recent years, post-editing of MT output is starting to become a common practice among human translators. However, it is well known that the quality of a given MT system can vary significantly across translation segments and that post-editing bad quality translations is a tedious task that may require more effort than translating texts from scratch. Previous research dedicated to learning quality estimation models to flag such segments has shown that models based on human annotation achieve more promising results. However, it is not yet clear what is the most appropriate form of human annotation for building such models. We experiment with models based on three annotation types (post-editing time, post-editing distance and post-editing effort scores) and show that estimations resulting from using post-editing *time*, a simple and objective annotation, can reliably indicate translation post-editing effort in a practical, task-based scenario. We also discuss some perspectives on the effectiveness, reliability and cost of each type of annotation.

1 Introduction

Post-editing Machine Translation (MT) output is now seen as a potentially successful way of incorporating MT into the human translation workflow in order to minimize time and costs in the translation industry. This is particularly true with Statistical MT (SMT) systems, which can be built with little effort from translation memories. However, a common complaint from human translators is that the post-editing of certain segments with low quality can be frustrating and can require more effort than translating those segments from scratch, without the aid of an MT system. Identifying such segments and filtering them out from the post-editing task is a problem addressed in the field of "Confidence Estimation" (CE), also called "Quality Estimation", for MT.

CE metrics are usually prediction models induced from data using standard machine learning algorithms fed with examples of source and translation features, as well as some form of annotation on the quality of the translations. Early work on sentence-level CE use annotations derived from automatic MT evaluation metrics (Blatz et al., 2004) such as BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and WER (Tillmann et al., 1997) at training time. The resulting models have not been shown to be effective, since the automatic metrics used do not correlate well with human judgments at the segment level and are difficult to interpret as absolute indicators of quality. More recent work focuses on having humans assigning absolute quality scores to translations, which has shown more promising results (Quirk, 2004; Specia et al., 2009a).

Obtaining explicit human annotations for translation quality, i.e., absolute scores reflecting post-editing effort, can however be a time-consuming and subjective task, requiring well trained annotators. In this paper we contrast prediction models learnt based on this type of annotation against two simpler and more objective variations of response variables: post-editing time and edit distance between automatic and post-edited transla-

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 73–80
Leuven, Belgium, May 2011

tions. We study how these response variables affect the task of CE in a practical scenario. For each CE model, besides computing error and correlation metrics with respect to human scores, we measure the actual post-editing time for unseen translations predicted as "good quality" according to such a model. We also compare these time measurements against that of post-editing all translations, without any filtering. This is the first attempt towards contrasting different types of human annotations and showing that using CE models that are good predictors of sentence-level post-editing effort to select a subset of translations for post-editing can speed up translation post-editing tasks.

In the remainder of this paper we first describe previous work on CE (Section 2), to then present the process of building datasets for two language pairs with the alternative human annotations (Section 3), the CE framework used in the experiments (Section 4), and our experiments and results in terms of standard metrics (Section 5) and within a task-based evaluation (Section 5.2). We conclude by discussing perspectives on the effectiveness, reliability and cost of each type of annotation types used the experiments, as well as some future work (Section 6).

2 Related Work

Blatz et al. (2004) present a number of experiments with CE at the sentence level based on annotations using automatic MT evaluation metrics. Regressors and classifiers are trained on features extracted for translations labeled according to NIST and WER. For classification, these scores are chosen to be thresholded to label the 5th or 30th percentile of the examples as "good". For regression, the estimated scores are mapped into two classes using the same thresholds. The results have not been found to be helpful in a range of evaluation tasks. This may be due to the fact that the automatic metrics used do not correlate well with human judgments. It may be also the case that the translations produced by the SMT systems at that time were too homogeneous in terms of quality: most translations would probably have been considered of bad quality by humans.

Quirk (2004) uses classifiers and a pre-defined threshold for "bad" and "good" translations considering a small set of 350 translations manually labeled for quality. Models trained on this dataset outperform those trained on a larger set of automatically labeled data. This provided a first indication that human annotation is much more effective for CE.

Specia et al. (2009a) use a number of "black-box" (MT system-independent) and "glass-box" (MT system-dependent) features to train a regression algorithm to estimate both NIST and human scores. While satisfactory accuracies were achieved with human annotations, the use of the estimated scores in a practical application was not tested.

He et al. (2010) use CE to recommend, for each source segment, a translation from either an MT system or a Translation Memory (TM) system for post-editing. Translation Edit Rate (TER) (Snover et al., 2006) is used to measure the distance between a reference translation (produced independently from the MT/TM systems) and each of these systems' output. At training time, this information is used to annotate sentences with a binary score indicating the system with the lowest TER (MT or TM). Based on a number of standard CE features, a classifier is trained to recommend the MT or TM for each new source segment. Therefore, TER is not directly used as an indicator of post-editing effort and is computed using references translations. Specia and Farzindar (2010) computed TER in a different way: between machine translations and their post-edited versions (HTER). These scores were then used to train a regression algorithm with standard CE features. While promising results were found in terms of correlation with human scores, they were not compared to models using any other form of human annotation. Evaluations with real applications to show the usefulness of the predicted scores were not performed.

Soricut and Echihabi (2010) focus on document-level CE. The goal is to rank the documents according to their estimated quality and, given a threshold defined by the end-user, select the top n documents for publishing. These are seen as documents whose automatic translation can be "trusted" as good enough for publishing, while the remaining documents are seen as not feasible for machine translation. Document-level CE constitutes a different problem, requiring different features and types of annotations (in their case, BLEU is used). Nevertheless, the view of CE as a ranking task to decide which texts are suitable for MT is an interesting one, which we also exploit in this paper.

3 Datasets with Human Annotations

The only datasets available from previous work on CE with human annotation focus on a single type of annotation (Specia et al., 2010). In this paper we present new datasets with the three annotation variants. These datasets were collected using *news* source sentences from development and test sets provided in different years of the WMT (Callison-Burch et al., 2010) evaluation campaign for two language-pairs, with variable sizes. The Moses toolkit (Koehn et al., 2007) was used to build SMT systems to produce translations for these source texts, based on the corpora and guidelines for the *baseline system* in WMT[1]:

- **fr-en** news-test2009: 2,525 French news sentences and their Moses translations into English (corpus-level BLEU = 0.2447).
- **en-es** news-test2010: first 1,000 English news sentences and their Moses translations into Spanish (corpus-level BLEU = 0.2830).

In order to gather human annotations in a way that is most natural to translators, a post-editing tool was built with a graphical interface similar to translation memory tools commonly used in the translation industry. The tool shows the source sentence and the translation produced by the MT system for post-editing. Post-editing time is measured on a sentence-basis in a transparent and controlled way, in order to isolate factors such as pauses between sentences.

Translators received initial training on the tool and task and were instructed to perform the minimum number of editions necessary to make the translation ready for publishing. They were aware of the time measurement and its general purpose. Within the tool, after post-editing each sentence, translators were asked to score the original translation according to its post-editing effort using options proposed in previous work (Specia et al., 2009a):

- 1 = requires complete retranslation.
- 2 = requires some retranslation, but post editing still quicker than retranslation.
- 3 = very little post editing needed.
- 4 = fit for purpose.

[1] http://www.statmt.org/wmt10/baseline.html

After the processing of each sentence, the edit distance between the original automatic translation and its post-edited version was computed using *Human Translation Edit Rate* (HTER) (Snover et al., 2006). HTER tries to estimate the number of edits that a human needs to perform in order to change the MT output into a good translation. Recent developments of the metric allow for matching of synonyms and paraphrases (Snover et al., 2010). However, we use standard HTER, which looks for exact matches only, since the post-edited translations here are expected to be as close as possible to the MT output, with only real errors corrected.

Edits in HTER include standard insertion, deletion and substitution of single words, as well as the shifting of word sequences. We set HTER options to tokenize the text, ignore case and use equal cost for all edits:

$$\text{HTER} = \frac{\#edits}{\#postedited\_words}$$

Even though the translators used in this paper were trained in the same way, both translation and quality annotations are subjective tasks and as a consequence certain measurements may vary considerably from translator to translator. This is particularly true with the measurement of *time*. Normalizing the annotations to account for such a variation is not straightforward. Moreover, these variations are natural and expected. Therefore, we believe that a CE model should be trained for each translator, based solely on their own annotation. In this paper, in order to guarantee consistency within datasets, each dataset was annotated by a single translator and models are built independently for each dataset.

The two translators who performed the task have different profiles. They both have a first degree in Translation Studies. However, the **fr-en** translator is a bilingual native speaker of English and French with very little professional experience in translation, while the **en-es** translator is a native speaker of Spanish and fluent speaker of English with considerable experience in translation tasks. None of the translators had experience with post-editing tasks. In discussions after the annotation task, it was clear that the two translators followed different post-editing strategies: the **en-es** translator resorted to external resources, such as bilingual dictionaries and concordancers, more often than the **fr-en** translator. This was mainly due to

the difference in the level of expertise with their respective language pair, but may also reflect a more careful approach taken by the **en-es** translator drawn from her experience with previous translation tasks. Once again, this emphasizes the need for CE models built for specific translators.

The annotation process resulted in three types of sentence-level annotation for each dataset:

1. Post-editing distance: a continuous score in $[0, 1]$, henceforth: *HTER*.
2. Post-editing effort score: a discrete integer score in $\{1,2,3,4\}$, henceforth: *effort*.
3. Post-editing time: average number of seconds to post-edit each word in the sentence, that is, number of seconds to post-edit the sentence normalized by the number of words in that sentence, henceforth: *time*.

The two datasets with these three types of annotations can be downloaded from `http://pers-www.wlv.ac.uk/~in1316/resources/datasets_ce_eamt.tar.gz` and can be used for training other CE models, tuning MT evaluation metrics, etc.

4 Confidence Estimation Framework

The CE framework used in this paper is similar to that proposed in (Specia and Farzindar, 2010): a Support Vector Machines regression algorithm with radial basis function kernel from the LIBSVM package (Chang and Lin, 2001)[2] and 80 shallow and MT system-independent features extracted from the source sentences and their corresponding translations, and also monolingual and parallel corpora. The features include:

- source & translation sentence lengths
- source & target sentence type/token ratio
- average source word length
- average number of occurrences of all target words within the target sentence
- source & translation sentence 3-gram language model probabilities obtained based on the source or target sides of the parallel corpus used to build the translation model of the SMT system
- translation sentence 3-gram language model probability trained on a POS-tags version of the target side of the parallel corpus used to train the SMT system
- percentage of 1 to 3-grams in the source sentence belonging to each frequency quartile of the source side of the parallel corpus used to train the SMT system
- average number of translations per source sentence word, as given by probabilistic dictionaries produced by GIZA++ (Och and Ney, 2003) trained on the same parallel corpus used to build the translation model of the SMT system
- percentages of numbers, content- / non-content words in the source & translation sentences
- number of mismatching opening/closing brackets and quotation marks in the translation sentence
- percentages of mismatches of superficial constructions between the source and translation sentences such as brackets, numbers, punctuation symbols, etc.

[2]With the parameters γ, ϵ and *cost* optimized.

5 Experiments and Results

Given the datasets annotated as described in Section 3 and the features described in Section 4, we trained three CE models for each language pair using a random subset of 90% of the source-translation sentence pairs. We then tested the models on the remaining sentences to compute standard error and correlation metrics, although these are not the focus of the evaluation task in this paper, as we discuss in Section 5.2. This procedure was repeated five times with different random samples for training and test. For correlation analysis, we use Spearman's rank coefficient, since for this task the ranking of the predicted scores is more relevant than their absolute values (see Section 5.2). The regression error is measured using Root Mean Squared Error (RMSE), which quantifies the average deviation of the estimated score with respect to the expected score:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}$$

where N is the number of test sentences, $\widehat{y}$ is the score predicted by the CE model and y is the actual score for that test sentence.

The results are shown in Table 1. The prediction errors are not comparable across annotation types, since these have different ranges of values. The comparison across language pairs is not straightforward, as the datasets have different sizes and were annotated by different translators. Nevertheless, these figures serve as basic indicators of the performance of the CE models. For example, taking the RMSE for datasets annotated with post-editing *effort*, one can see that on average the models do not make mistakes that cross more than one of the four categories. The average error of this *effort* model is smaller in the **en-es** datasets, despite the fact that it uses significantly smaller training sets.

The average error of *HTER* models is comparable in both datasets. The average error of models trained using *time* is considerably higher in the smaller dataset: $\sim$2 seconds per word. As discussed in Section 3, this can be due to the much larger time variation in the post-editings made by the **en-es** translator: 0-54 seconds per word, while the **fr-en** translator spent 0-10 seconds per word. This shows that translators perform post-editing in different ways based on the language pair, their experience, knowledge of the domain, etc., even for similar texts. For this particular task, both **fr-en** and **en-es** datasets were very similar in nature: same genre and source (news), similar time periods (2009-2010), similar average sentence length (22 words).

Dataset		RMSE	Spearman
fr-en	*HTER*	0.155 ± 0.011	0.366 ± 0.047
	effort	0.662 ± 0.022	0.459 ± 0.034
	time	0.651 ± 0.040	0.455 ± 0.052
en-es	*HTER*	0.178 ± 0.006	0.281 ± 0.102
	effort	0.549 ± 0.028	0.367 ± 0.096
	time	1.970 ± 0.250	0.298 ± 0.024

Table 1: Average error and Spearman's correlation coefficient in the test sets labeled with different types of human annotation

The correlation analysis can provide a better basis for comparing models built with different annotation types. In both datasets, the best correlation score was obtained with the *effort* models, although the *time* model achieved very similar correlation in the **fr-en** datasets and was much more stable than the *HTER* model in the **en-es** datasets (smaller standard deviation). Overall, the correlation scores are higher for **fr-en** models, which may be due to the fact that they are built using twice as many the training examples as the **en-es** models.

5.1 Feature Analysis

In order to investigate the contribution of different features to the CE models, as well as compare the relevance of features across all three type of prediction tasks and the two language pairs, we checked Pearson's correlation coefficient of each individual feature and the specific type of annotation. The three best correlated features for each of these variations are shown in Tables 2 and 3.

Dataset	Top 3 Features
HTER	1. average number of translations per source word in the sentence weighted by the inverse frequency of each word in the source corpus 2. ratio of source by target lengths 3. ratio of percentage of tokens a-z in the source and target sentences
effort	1. 3-gram LM score - source sentence 2. 3-gram LM score - target sentence 3. 3-gram LM score - target POS tags
time	1. 3-gram LM score - source sentence 2. 3-gram LM score - target sentence 3. 3-gram LM score - target POS tags

Table 2: Correlation analysis of individual features in the **fr-en** datasets

The top ranked features are the same or similar for *effort* and *time* annotations. Moreover, with these annotations, the correlation of certain individual features is clearly higher than that of most other features. With the *HTER* annotations, however, the correlations obtained with most features are very close to each other and relatively lower than those obtained with the other annotation types.

While more sophisticated feature analysis can be done, such as building models with one feature at a time and all except one feature at a time, this simple analysis already shows that features based on language models of the source and target sentences and average ambiguity of the source words (given by probabilistic dictionaries) perform the best.

Dataset	Top 3 Features
HTER	1. average number of translations per source word in the sentence weighted by the inverse frequency of each word in the source corpus 2. average number of translations per source word in the sentence 3. average source unigram frequency in the 4th quartile of frequency
effort	1. 3-gram LM score - source sentence 2. 3-gram LM score - target POS tags 3. lenght of source/target sentence
time	1. 3-gram LM score - source sentence 2. % of nouns - source sentence 3. average number of translations per source word in the sentence weighted by the inverse frequency of each word in the source corpus

Table 3: Correlation analysis of individual features in the **en-es** datasets

5.2 Task-based Evaluation of CE Models

Although error and correlation scores give some indication of the performance of CE models, they do not serve to assess the actual benefit from using the CE models in a translation post-editing task. In order to assess the actual effectiveness of the CE models learnt from different annotation types, we propose measuring the number of words that can be post-edited in a fixed amount of time in translations selected according to each resulting CE model.

For this *task-based* evaluation, unseen sentences with the same genre and domain as those used to build the CE models were selected from different WMT releases. They were translated using the same SMT systems described in Section 3:

- **fr-en** news-test2010: 2,489 French news sentences and their Moses translations into English (corpus-level BLEU = 0.2551).
- **en-es** news-test2009: 2,525 English news sentences and their Moses translations into Spanish (corpus-level BLEU = 0.2428).

Quality predictions were generated for these test sets using the three variations of the CE models. The predicted scores can be exploited in different ways, ranging from simply giving them to translators to inform them of the estimated quality to using them to directly filter out bad quality translations from the post-editing workflow. We believe the best way to use the predicted scores is to directly select a subset of machine translations with (supposedly) good quality for post-editing, while giving the remaining cases for translation. Alternative ways of setting a threshold on the estimated scores to create such a subset can be found in (Specia et al., 2009b) and (He et al., 2010). In this paper we are rather interested in the assessment of different types of annotations. Therefore, we evaluate the ranking of translations using alternative CE models. Our focus is on checking whether the CE scores can be beneficial in post-editing tasks. More specifically, we would like to answer the following questions:

1. Which annotation type yields models that allow ranking sentences in a way that selecting the best ranked sentences can maximize the number of words that can be post-edited per second?
2. Using CE models to rank sentences and selecting only a subset of the best ranked sentences, is it possible to post-edit more words as compared to post-editing sentences without any ranking in a given slot of time?

In order to answer these questions, we randomly selected four subsets of 600 translations from each unseen dataset. The translations in three of these subsets were then ranked using each CE model so that the (supposedly) best translations appear first. Translations in the forth set were not ranked. The size of the subsets guarantees enough variation in the quality of the translations. In fact, since these unseen texts are similar to the ones used to train the models, we expect the translations to follow a similar distribution in terms of quality scores, which is not very skewed. For example, if we take the *effort* annotation for the **en-es** datasets, which has a more straightforward interpretation, approximately 43% of the translations were considered "good" (scores 4 or 3) while the renaming 57% were considered "bad" (scores 1 or 2).

The same translators who performed the initial annotation were asked to post-edit as many sentences as possible following their order in four "tasks" on different days, dedicating **one-hour per task** and using the same annotation tool (without the scoring facility). The order of the tasks was randomly defined:

- T1: 600 machine translated sentences sorted according to the *HTER* model.
- T2: 600 machine translated sentences sorted according to the *effort* model.
- T3: 600 machine translated sentences sorted according to the *time* model.
- T4: 600 machine translated sentences without any sorting.

For each dataset, Table 4 shows the average number of words post-edited per second, along with the number of sentences post-edited per hour (notice that sentences have variable sizes). The latter figure refers to the total number of words in the final post-edited sentences, including words which were kept as in the original MT. As an upper bound, if the original machine translations were perfect, no post-edition would need to be made, and the time spent with the post-editing task would only include the reading of the source and translation sentences.

Dataset		Sentences/h	Words/s
fr-en	T1: *HTER*	65	0.96
	T2: *effort*	97	0.91
	T3: *time*	82	1.09
	T4: unsorted	55	0.75
en-es	T1: *HTER*	38	0.41
	T2: *effort*	71	0.43
	T3: *time*	69	0.57
	T4: unsorted	33	0.32

Table 4: Results of the task-based evaluation: number of sentences ranked according to different CE models that can be post-edited in one hour, as well as the corresponding number of words that can be post-edited per second.

For both language pairs, post-editing only the top machine translations according to any CE model allows more words to be post-edited per second than post-editing any machine translations ("unsorted"). The best rate is obtained with *time* as response variable in both **fr-en** and **en-es** datasets, contrary to what was found using the correlation metric (Table 1) for the **en-es** dataset, showing the value of this task-based evaluation.

Overall, these results show that the explicit and subjective type of annotation used in previous work, post-editing *effort*, is not better than simpler and more objective metrics: *time* and *HTER*, which can be both obtained as a by-product of having humans post-editing a reasonably small number of machine translations. In particular, using *time*, which is a very intuitive and transparent way of measuring post-editing effort, clearly outperforms all other types of annotations.

Although in real-world scenarios translators/post-editors would have to translate a complete set of sentences, as opposed to the top ranked sentences according to a CE model, the idea is that a reliable CE model could help distinguishing sentences that are worth post-editing from those that should be translated from scratch. This could not only increase productivity by preventing translators from spending time reading bad quality translations that are not worth post-editing, but also to avoid translators' frustration with trying to post-edit bad quality translations.

6 Conclusions and Future Work

We have presented experiments with alternative ways of annotating translation quality for building confidence estimation models for MT. The results in a practical, task-based evaluation show that CE models learnt from objective annotations of translation quality produce rankings of translations that reliably reflect their post-editing effort. This in turn can be used to minimize post-editing time and, more important, the frustration that human translators may feel when asked to post-edit bad quality translations.

While in general it is recommended that a CE model is built for each language pair, MT system and human translator, collecting a reasonably small number of post-editings is sufficient to build such models. Considering a translation workflow where professional translators already post-edit the output of MT systems, post-editing *time* and *HTER* annotations can be obtained in a transparent and cost-effective way with simple post-editing tools like the one used here. This offers a great advantage as compared to the expensive, time consuming and subjective task of asking human annotators to explicitly judge translations according to their quality.

In future work, we plan to perform similar experiments using ranking algorithms, as opposed to regressors, as well as combine these algorithms with techniques to establish thresholds on the predicted scores. In addition, we will seek to design

a post-editing tool that can incorporate CE predictions for translations coming from one or more translation tools in a seamless and transparent way to translators.

We also plan to use crowdsourcing mechanisms such as Mechanical Turks to include other datasets in our study, as well as to ensure the quality of the post-editing by including multiple post-editors and reviewers for each dataset. We would also like to analyze changes in the behavior of translators as they gain more experience with the task of post-editing, especially with respect to the annotations using post-editing time.

References

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 the 20th Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden.

Chang, Chih-Chung and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Doddington, George. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technology*, pages 138–145, San Diego, California.

He, Yifan, Yanjun Ma, Josef van Genabith, and Andy Way. 2010. Bridging SMT and TM with Translation Recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 622–630, Uppsala, Sweden.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics: Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

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

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

Quirk, Chris. 2004. Training a Sentence-Level Machine Translation Confidence Measure. In *Proceedings of the 4th Conference on Language Resources and Evaluation*, pages 825–828, Lisbon, Portugal.

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 the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231.

Snover, Matthew, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2010. TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3):117–127.

Soricut, Radu and Abdessamad Echihabi. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden.

Specia, Lucia and Atefeh Farzindar. 2010. Estimating machine translation post-editing effort with hter. In *Proceedings of the AMTA-2010 Workshop Bringing MT to the User: MT Research and the Translation Industry*, Denver, Colorado.

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

Specia, Lucia, Marco Turchi, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders. 2009b. Improving the confidence of machine translation quality estimates. In *Proceedings of the Machine Translation Summit XII*, Ottawa, Canada.

Specia, Lucia, Nicola Cancedda, and Marc Dymetman. 2010. A dataset for assessing machine translation evaluation metrics. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Tillmann, Christoph, Stefan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated DP based Search for Statistical Translation. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 2667–2670, Rhodes, Greece.