Estimating Machine Translation Post-Editing Effort with HTER

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

Atefeh Farzindar
NLP Technologies Inc. and
University of Montreal
Montreal, Canada
farzindar@nlptechnologies.ca

Abstract

Although Machine Translation (MT) has been attracting more and more attention from the translation industry, the quality of current MT systems still requires humans to post-edit translations to ensure their quality. The time necessary to post-edit bad quality translations can be the same or even longer than that of translating without an MT system. It is well known, however, that the quality of an MT system is generally not homogeneous across all translated segments. In order to make MT more useful to the translation industry, it is therefore crucial to have a mechanism to judge MT quality at the segment level to prevent bad quality translations from being post-edited within the translation workflow. We describe an approach to estimate translation post-editing effort at sentence level in terms of Human-targeted Translation Edit Rate (HTER) based on a number of features reflecting the difficulty of translating the source sentence and discrepancies between the source and translation sentences. HTER is a simple metric and obtaining HTER annotated data can be made part of the translation workflow. We show that this approach is more reliable at filtering out bad translations than other simple criteria commonly used in the translation industry, such as sentence length.

1 Introduction

One of the most popular ways to incorporate Machine Translation (MT) into the translation workflow is to have humans checking and post-editing the output of MT systems. This is the procedure followed, for example, by NLP Technologies[1], which has developed a translation management system called TRANSLI^{TM}specifically tailored for the legal field and based on Statistical Machine Translation (SMT) followed by post-edition by expert reviewers. However, the post-editing of a proportion of the translated segments may require more effort than translating those segments from scratch, without the aid of an MT system. This problem has been addressed with metrics of "Confidence Estimation" (CE) for MT.

CE metrics use features extracted from the machine translations, and usually also from the source text and monolingual and bilingual corpora, and optionally information about the MT system used to produce the translations. Such features are given to a machine learning algorithm in order to learn a model to predict quality estimates for a certain language pair, MT system and text domain/genre from data annotated with scores reflecting translation quality derived either from automatic MT evaluation metrics (Blatz et al., 2004) such as NIST (Doddington, 2002) and WER (Tillmann et al., 1997) or using human annotation (Quirk, 2004; Specia et al., 2009a). CE metrics may provide a score to end-users of MT systems for each word or phrase (Gandrabur and Foster, 2003; Ueffing and Ney, 2005; Kadri and Nie, 2006), sentence (see Section 2) or document (Soricut and Echihabi, 2010) translated by the MT system. This paper focuses on sentence-level confidence estimation.

Early work has focused on binary indicators of translation quality to filter out bad translations from

[1] http://www.nlptechnologies.ca

Ventsislav Zhechev (ed.): *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" (JEC '10)*, pp. 33–41. Denver, CO, 4 November 2010.

being post-edited by professional translators (Blatz et al., 2004; Quirk, 2004). More recent work focuses on estimating a continuous numeric score that can be used directly to inform human translators of translation quality (Specia et al., 2009a) or thresholded according to the requirements of a given task and level of experience of the professional translator (Specia et al., 2009b).

Using human scores to train CE systems has been shown to be more effective than using automatic MT evaluation metric scores such as NIST or WER (Quirk, 2004), which is expected, given that such metrics do not correlate very well with human judgments at the segment level. However, producing human annotation is a time-consuming and subjective task, and unless annotators are well trained for the task of assigning absolute quality scores to translations, the scores obtained may be inconsistent and therefore not adequate to train machine learning algorithms. In this paper we exploit a simpler, cheaper and more objective type of score, produced by a semi-automatic MT evaluation metric that is known to correlate well with human judgments at the segment level: Human-targeted Translation Edit Rate (HTER) (Snover et al., 2006). This metric consists in measuring the edit distance between the translation produced by the MT system and its minimally post-edited version produced by a human translator. Our learning framework is therefore trained on HTER scores to directly estimate translation post-editing effort in terms of this metric. We show that this metric is more reliable in filtering out bad translations than other simple criteria commonly used in the translation industry, such as sentence length.

In the remainder of the paper we first refer to related work on CE (Section 2), then describe our experimental setup (Section 3) and report the results of a number of experiments (Section 4).

2 Related Work

The first comprehensive investigation on CE at the sentence level is that of Blatz et al. (2004). Regressors and classifiers are trained on features extracted for translations labeled according to MT metrics like NIST. For classification, NIST 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 did not show to be helpful to the tasks evaluated, which may be due to the automatic metrics used.

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.

Specia et al. (2009a) use a number of "black-box" (MT system-independent) and "glass-box" (MT system-dependent) features to train Partial Least Squares (PLS) regression to estimate both NIST and human scores. While satisfactory accuracy was achieved with human annotations, the use of the estimated scores in a practical application was not tested. In (Specia et al., 2009b), the technique of Inductive Confidence Machines was used to allow the automatic identification of a threshold to map a continuous predicted score (based on human annotation) into good / bad categories for filtering out bad-quality translations. This threshold is defined according to the expected confidence level of the system. The application of the estimated confidence scores to filter out bad sentences or select the best translation from multiple MT systems is presented in (Specia et al., 2010b). While promising results were found for both applications, the approach is dependent on human annotation to train the system.

He et al. (2010) use CE to recommend a translation from either an MT system or a Translation Memory (TM) system for post-editing. Standard translation Edit Rate (TER) is used to measure the distance between a reference translation (produced independently from the MT/TM systems) and each of these systems' output. This information is used to annotate source sentences with a binary score to indicate its lowest TER (MT or TM) and train a classifier to recommend the translation aid tool most likely to be useful for a new source sentence. Therefore, TER is not directly used as an indicator of post-editing effort.

3 Experimental Setup

In this paper we experiment with using an automatic MT evaluation metric to score translations which is

very objective and easy to obtain as a by-product of the use of MT systems by professional translators, but correlates well with human absolute judgments on translation quality. As we present in what follows, a corpus of legal documents is first translated using a given MT system, and then post-edited by human translators. A modified version of the edit distance is then measured between the machine translation and its post-edited version. This distance is used to train an implementation of Support Vector Machines for regression. Standard evaluation metrics are then computed on a held-out test set, and the produced scores are used in various applications.

3.1 Annotation and Evaluation Metrics

Translation Edit Rate (HTER) In order to annotate each sentence for translation quality, we use HTER (Snover et al., 2006) to measure the distance between the MT output and its post-edited version. HTER measures the amount of editing that a human would have to perform to change the MT output to make it a good translation. Therefore, the human post-edited version is considered here the "reference" translation.

Recent developments of the metric allow for matching of synonyms and paraphrases (Snover et al., 2010), however we use the standard TER which looks for exact matches only, since the "reference" translations here are more likely to be the closest possible to the MT output, where only real mistakes are corrected. HTER is defined as the minimum number of edits needed to change the MT output so that it matches exactly the reference, normalized by the length of the reference. Edits include insertion, deletion and substitution of single words, as any standard edit distance metric, as well as shifts of word sequences. In the version we used[2], the text is pre-tokenized and all edits are case sensitive and have equal cost:

$$\text{TER} = \frac{\#edits}{\#reference\_words}$$

As opposed to the expensive, time consuming and subjective task of asking human annotators to explicitly judge translations according to their quality, the process of obtaining annotations according to HTER can be incorporated into the translation workflow in a simple and cost-effective way. Considering a translation workflow where professional translators already post-edit the output of MT systems, one only needs to instruct the translators to produce the minimum editing necessary to make the translations publishable, collect a reasonably small number of translations (as we will discuss in Section 4) and then apply the HTER metric to automatically measure the distance between the original translation and its post-edited version.

Root Mean Squared Prediction Error (RMSPE) In order to evaluate the performance of the CE system, we compute the average error in the estimation of TER scores using the RMSPE metric:

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

where N is the number of test sentences, $\widehat{y}$ is the HTER/TER predicted by the learning algorithm and y is the actual value of the HTER/TER for that test case. RMSPE quantifies the average deviation of the estimator with respect to the expected score: the lower the value, the better the performance of the CE system.

Pearson's Correlation To evaluate the performance of the CE system, we compute Pearson's correlation coefficient between the predicted score $\widehat{y}$ and the expected score y. Pearson's correlation coefficient between the expected and predicted scores measures their linear dependence and is defined as the covariance of these two variables divided by the product of their standard deviations:

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

This metric is traditionally used for the assessment of machine translation evaluation metrics. It allows analyzing whether the difference between the predicted and expected scores is or is not simply a matter of re-scaling the predictions. The higher its

[2]The command *terp_ter* available from `http://www.umiacs.umd.edu/~snover/terp/`, with the options -c to cap the metric so that it varies within [0, 1], and with or without -s for case insensitive or sensitive variations.

absolute value, the better the performance of the CE system.

3.2 Datasets

Fr-En Legal Dataset Our main dataset consists of English translations of French legal documents. Legal documents are produced in large quantities and if they are related to a bilingual or multilingual country/community, they need to be quickly and accurately translated into all its official languages. Our dataset was generated by TRANSLI^{TM}, the certified translation system and service provided by NLP Technologies. TRANSLI^{TM}consists of an SMT system similar to Portage (Johnson et al., 2006) trained on legal documentation to produce initial translations, which are then manually post-edited by experts in the legal domain. Details about the corpus can be found in (Farzindar, 2009)

We have put together all 25 documents translated, segmented them into sentences and filtered out those sentences which were not translations of each other (true transcriptions). This process resulted in $2,349$ sentences in each language. We then annotated each sentence for translation quality using HTER, as described in Section 3.1, to produce $2,349$ quadruples of the type:

$$\{source, translation, post-edition, hter\_score\}$$

The dataset was randomly split into training (85%) and test (15%) using a uniform distribution.

En-Es Europarl Dataset Additionally, we used four datasets available in (Specia et al., 2010a). Each dataset consists of $4,000$ Spanish translations for English sentences taken from the Europarl development and test sets provided by WMT08 (Callison-Burch et al., 2008), produced by four Statistical MT (SMT) systems trained on Europarl (S1-S4). To compute TER, we used the reference translations provided by WMT08, which are translations manually produced by humans, without any MT system, as opposed to post-editions of MT output. These previously produced translations are known to be less adequate for edit distance metrics, since the fact that the human and machine translations are different does not necessarily indicate that the machine translations are incorrect. Human annotations for quality are also available for these translations. Professional translators assigned a quality score in $[1-4]$ to each translation, which is a range commonly used by them to indicate the quality of translations with respect to the need for post-editing, as we will discuss in Section 4.4. The resulting datasets consist of four sets of $4,000$ quintuples of the type:

$$\{source, translation, reference,$$

$$ter\_score, human\_score\}$$

Each dataset was randomly split into training (75%) and test (25%) using a uniform distribution.

3.3 Features

We use 86 standard shallow features extracted from the input (source) sentences and their corresponding translation (target) sentences, and also monolingual and parallel corpora:

- source & target sentence lengths and their ratio;
- source & target sentence type/token ratio;
- average source word length;
- source & target sentence unigrams, bigrams and trigram language model probabilities and perplexities obtained using the source/target side of the corpus used to train the SMT system as monolingual corpus;
- target sentence trigram language model probability trained on a POS-tagged corpus of the target language (Europarl);
- average frequency of unigrams, bigrams and trigrams in the source sentence belonging to each frequency quartile of a corpus of the source language (Europarl);
- average frequency of source sentence unigrams in a source language corpus (Europarl);
- percentage of unigrams, bigrams and trigrams in the source sentence belonging to each frequency quartile of a corpus of the source language (Europarl);
- percentage of distinct unigrams, bigrams and trigrams in the source sentence seen a corpus of the source (Europarl);

- average number of translations per source word in the sentence, as given by probabilistic dictionaries produced by GIZA++ (Och and Ney, 2003) extracted from the parallel corpus used to train the SMT system, thresholded using different percentages (0.01, 0.05, 0.10, 0.20, 0.50), unweighted or weighted by the direct or inverse frequency of the words in the source language corpus;
- percentages of punctuation symbols, numbers, content- / non-content words in the source & target sentences and their ratio;
- number of mismatching opening/closing brackets in the target sentence;
- whether target sentence contains mismatched quotation marks;
- number of mismatches of each of the following superficial constructions between the source and target sentences: brackets, each punctuation symbol (and all of them together), numbers, either in absolute terms or normalized by sentence length; and
- proportion of words in the source and target with initial/all/none capital letters, or only capital letters and symbols, and the ratio between the proportions of words in the source and target sentences with such case patterns[3].

3.4 Learning Algorithm

We use an implementation of Support Vector Machines (SVM) for regression: epsilon-SVR algorithm with radial basis function kernel from the LIBSVM package (Chang and Lin, 2001), with the parameters γ, ϵ and *cost* optimized.

The optimization of the SVM parameters in both types of datasets was performed by cross-validation using five random subsamples of the training set (75% for validation training and 25% for validation test).

4 Results

In what follows we describe a number of experiments with our main dataset, **Fr-En Legal**, as well as the **En-Es Europarl** datasets for comparison.

[3]These features can only be used with the **Fr-En Legal** dataset, as the other datasets are not available in their truecase version.

4.1 Prediction Error and Correlation Scores

Table 1 shows the prediction error and correlation scores obtained with our five datasets annotated with HTER/TER basic version, that is, exact match and identical weights to all edit operations.

Dataset	RMSPE	Pearson
Fr-En Legal	0.1683	0.6292
En-Es Europarl S1	0.1777	0.3351
En-Es Europarl S2	0.1732	0.3870
En-Es Europarl S3	0.1657	0.3583
En-Es Europarl S4	0.1402	0.3979

Table 1: RMSPE and Pearson's correlation scores obtained for each datasets annotated with HTER/TER. Both prediction error and correlation are computed against the HTER/TER annotation of the test set.

RMSPE indicates that, on average, the predicted HTER/TER score deviates from the true HTER/TER score from 0.14 to 0.18. While this might appear to be a high deviation, considering that the metric varies between 0 and 1, it is difficult to measure its impact in the practical use of the predicted CE scores. For example, if all predicted scores are consistently lower or higher than the actual scores, this would mean that it is necessary to simply scale the predicted score accordingly. We therefore consider that correlation scores are more relevant, as they can show whether or not the differences are a matter of scaling the data. CE achieves a considerably higher correlation with the true HTER scores for the **Fr-En Legal** dataset. The lower performance of the **En-Es Europarl** datasets can be explained by the fact that TER in this case may not reflect post-editing effort appropriately, since it was computed using reference translations as opposed to post-edited translations.

In these experiments each dataset is first normalized through lowercasing and tokenization, and therefore edits due to punctuation or case are treated equally as incorrect word edits and the case-sensitive features are not used. We also performed experiments without normalizing the translations and their post-edited version for the **Fr-En Legal** dataset, while using explicit features to cover differences in the case patterns between translations and their post-editions. However, contrary to our expectations, the use of case sensitive options and features resulted

in lower performance. This could be because the case-dependent features reflect very simple statistics about case in source and translation sentences. More advanced features could use word alignment information to capture the differences appropriately. Nevertheless, we consider that editings due to case are very straightforward and cheap to be made, and should not thus be counted as a standard substitution, that is, as a possibly completely incorrect word. Therefore, only case insensitive results are reported this paper.

Although HTER/TER allows using options to identify synonyms or paraphrases between the reference and machine translations, we did not use this option in our experiments for different reasons. With the **Fr-En Legal** dataset, since the translators were instructed to perform the minimum post-editing necessary, they are very unlikely to have paraphrased the translation and therefore using paraphrasing resources could add noise to the HTER scores computed. On the other hand, these resources could be useful for the **En-Es Europarl** datasets, where paraphrases are widely found, but they are not available for translations into Spanish.

4.2 Filtering Out Bad Translations

In Table 2 we contrast the performance of our CE approach in predicting post-editing effort in terms of HTER/TER scores to other criteria commonly used to filter out potentially bad machine translations from post-editing: (1) the size of the input segment in words (*Size*), as it is usually believed that long segments are likely to be incorrectly translated ; and (2) a 3-gram language model score of the input segment using the source side of the SMT training corpus to compute the language model (*LM*), which can be seen as an approximation to the fuzzy match level metric of translation memories, since it is well known that common segments in the training corpus are likely to be translated correctly. We report Pearson's correlation between each of these criteria and the expected HTER/TER score.

The considerably lower performance of the *LM* and *Size* criteria for the **En-Es Europarl** datasets can be explained again by the fact that TER may not reflect post-editing effort appropriately in this case, since it was computed using reference translations.

In order to contrast the performance of our CE approach against that of the *LM* and *Size* criteria in a more intuitive way, we take the **Fr-En Legal** dataset and look into the use of these three criteria to filter out potentially bad translations. This could be done by establishing thresholds on the edit distance above which the machine translations should be filtered out from the translation workflow to save post-editors' time. We could then check which of the criteria are able to select a larger number of translations within the same threshold. However, establishing the exact threshold on HTER scores above which translations should be considered too bad to be post-edited is a complex problem in itself. We instead establish a percentage of translations to filter out. For example, if we assume that the worst 10% of the machine translations should be filtered out, we can look at each of our three criteria (CE score, source sentence LM score and source sentence size) to check whether the translations they indicate as the worst 10% agree with those pointed out by the true HTER score. In Table 3 we show the results for four thresholds: 10%, 25%, 50% and 65%, as well as the expected total of sentences to be selected according to the true HTER (out of 349 test sentences) in the second column.

Dataset	CE	LM	Size
Fr-En Legal	**0.6292**	0.2316	0.2768
En-Es Europarl S1	**0.3351**	0.1396	0.0875
En-Es Europarl S2	**0.3870**	0.1607	0.0848
En-Es Europarl S3	**0.3583**	0.1439	0.0600
En-Es Europarl S4	**0.3979**	0.1098	0.0819

Table 2: Pearson's correlation scores obtained for each dataset annotated with HTER/TER and our CE score, along with the correlation of HTER/TER and other criteria commonly used to filter out potentially bad translations.

As we can see, the CE score is able to correctly select the highest number of bad quality translations within all percentage thresholds. For large percentages to be filtered out, all criteria become similar, which is expected, given that larger portions of the test are covered by those percentage thresholds. The more strict the percentage, the more difficult it is for any criteria to select the correct bad quality translations, and the larger the difference between CE and other criteria, showing its advantage. Nevertheless,

Percentage	HTER	CE	Size	LM
10%	34	**12**	3	2
25%	84	**39**	26	22
50%	169	**133**	103	110
65%	226	**194**	185	182

Table 3: Number of sentences that should be filtered out (in HTER) according to different pre-defined percentages of "bad" translations, along with the number of sentences that would be correctly filtered out according to different filtering criteria: CE score, size and LM of the source sentence.

CE still only selects a subset of the total number of bad quality translations (30% to 86% of that total, depending on the percentage threshold). This may be an indication that we need to adjust the thresholds accordingly. For example, it may be necessary to filter out the bottom 20% translations scored according to the CE in order to be able to select the bottom 10% of the true HTER. This will of course have an influence on the number of good translations which are incorrectly filtered out. The choice will depend on the intended use of CE: filtering out most potentially bad quality translations at the cost of filtering out some good quality translations as well, or making sure the maximum number of good translations is kept for post-editing. A solution to tune the filtering threshold is proposed by (Specia et al., 2009b).

4.3 Effect of the Number of Training Instances

Obtaining annotated data for training a CE system using HTER is straightforward assuming that using an MT system followed by human post-editing is already part of the translation workflow, or that this scenario can be introduced into the translation workflow to gather some initial data. For a given language pair, text domain and genre and MT system, one can collect a number of triples including the source texts, their machine translations and post-edited versions, compute HTER to then train the machine learning algorithm. In this section we perform a small set of experiments to analyze the effect of the number of training instances on the performance of the CE approach with the **Fr-En Legal** dataset. The complete training set, which is already relatively small, with only 2,000 sentences, was randomly subsampled to select smaller numbers of training instances. A CE model was generated using each of these subsamples and the same test set as was used to evaluate the models. The results are shown in Table 4.

Training instances	Pearson
2,000	0.6292
1,500	0.6099
1,000	0.5646
500	0.4516

Table 4: Pearson's correlation scores obtained with different subsets of the training instances from the **Fr-En Legal** dataset and the same test set.

The exact number of training instances necessary to train a CE system will depend on features of the dataset, such as the distribution of quality scores in the dataset. In this particular case, no significant drop in the correlation is observed with 1,500 training cases. For 1,000 and smaller training sets, the correlation scores drop considerably.

4.4 TER and Human Scores

For the **En-Es Europarl** datasets, quality scores assigned by professional translators to each translation in a [1, 4] range are also available. The professional translators received training to score the translations using the following options:

- 1 = requires complete retranslation;
- 2 = post editing quicker than retranslation;
- 3 = little post editing needed; and
- 4 = fit for purpose.

We compare our predictions using SVM trained on TER scores against the predictions obtained using SVM trained on human annotation as reported by (Specia et al., 2010b). We contrasted the use of the predictions to filter out bad translation, as in Section 4.2. In our experiment, translations scored 2-4 should be kept for post-editing, while translations scored 1 ("requires complete retranslation") should be filtered out. As we show in Table 5, 29 translations were scored by humans with 1, corresponding to approximately 3% of the test set. If we consider the bottom 3% of translations as scored by the CE

approach trained on human scores, we are able to filter out 34% (10) of these bad quality translations. If we instead take the bottom 3% of translations as scored by the CE approach trained on TER scores, we are only able to filter out 21% (6) of these translations. These results are nevertheless better than those obtained using the source sentence size and language model criteria, which are able to filter out only 1 or 2 bad quality translations, respectively.

Human	CE+Human	CE+TER	Size	LM
29 (3%)	**10**	6	1	2

Table 5: Number of bad quality sentences (scored 1 according to humans) that would be correctly filtered out according to different CE approaches (CE+Human = CE training using human scores; CE+TER = CE trained using TER scores) and other filtering criteria (size and LM score of the source sentence).

It is important to recall that the edit distance metric used to annotate the **En-Es Europarl** datasets is TER, as opposed to HTER, which does not correlate as well with human judgments. Once again, improvements could be obtained by adjusting the thresholds according to the intended use of CE, as discussed in Section 4.2.

5 Conclusions

Having in mind a translation workflow where professional translators check and post-edit the output of MT systems when necessary, we have presented an approach to estimate the post-editing effort of translations produced by MT systems and save translators' time by preventing them from post-editing bad quality translations.

We have shown that, although having human annotations of post-editing effort to train our approach would be the ideal scenario, it is possible to obtain a good performance with simpler and cheaper annotations by collecting a small set of machine translations and their post-edited versions and computing HTER, a semi-automatic translation edit rate metric.

The approeach presented here will be integrated to the production MT system for English and French legal documents. We also hope to be able to exploit datasets for other language pairs, and particularly to better study whether we can approximate human annotation using HTER or other automatic metrics for this particular problem of CE.

Acknowledgments

We thank Michel Simard from the National Research Council Canada (NRC) for kindly providing us some of the data to extract the confidence estimation features.

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 *20th Coling*, pages 315–321, Geneva.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *3rd Workshop on Statistical Machine Translation*, pages 70–106, Columbus.

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.

Farzindar, Atefeh. 2009. Automatic Translation Management System for Legal Texts. In *MT Summit XII: Proceedings of the twelfth Machine Translation Summit*, pages 417–424, Ottawa, Ontario, aug.

Gandrabur, Simona and George Foster. 2003. Confidence estimation for translation prediction. In *7th Conference on Natural Language Learning*, pages 95–102, Edmonton.

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, July.

Johnson, Howard, F. Sadat, George Foster, Roland Kuhn, Michael Simard, Eric Joanis, and S. Larkin. 2006. Portage with Smoothed Phrase Tables and Segment Choice Models. In *Workshop on Statistical Machine Translation*, pages 134–137, New York.

Kadri, Youssef and Jian-Yun Nie. 2006. Improving query translation with confidence estimation for cross-language information retrieval. In *15th ACM International Conference on Information and Knowledge Management*, pages 818–819, Arlington.

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

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

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 (Special Issue on: Automated Metrics for MT Evaluation)*.

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, July.

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

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*, August.

Specia, Lucia, Nicola Cancedda, and Marc Dymetman. 2010a. A Dataset for Assessing Machine Translation Evaluation Metrics. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Specia, Lucia, Dhwaj Raj, and Marco Turchi. 2010b. Machine translation evaluation versus quality estimation. *Machine Translation*, pages 1–12.

Tillmann, C., S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated Dp Based Search For Statistical Translation. In *European Conference on Speech Communication and Technology*, pages 2667–2670.

Ueffing, Nicola and Hermann Ney. 2005. Application of Word-Level Confidence Measures in Interactive Statistical Machine Translation. In *10th Meeting of the European Association for Machine Translation*, pages 262–270, Budapest.