Predicting Machine Translation Adequacy

Lucia Specia, Najeh Hajlaoui, Catalina Hallett and **Wilker Aziz**
Research Group in Computational Linguistics
University of Wolverhampton
{l.specia, najeh.hajlaoui, c.hallett, w.aziz}@wlv.ac.uk

Abstract

As Machine Translation (MT) becomes more popular among end-users, an increasingly relevant issue is that of estimating the quality of automatic translations for a particular task. The main application for such quality estimates has been selecting good enough translations for human post-editing. The end-users, in this case, are fluent speakers of both source and target languages and the quality estimates reflect post-editing effort, for example, the time to post-edit a sentence. This paper focuses on quality estimation to address the challenging problem of making MT more reliable to a different type of end-user: those who cannot read the source language. We propose a number of indicators contrasting the source and translation texts to predict the adequacy of such translations at the sentence-level. Experiments with Arabic-English MT show that these indicators can yield improvements over previous work using general quality indicators based on source complexity and target fluency.

1 Introduction

The use of Machine Translation (MT), and particularly Statistical MT (SMT), is becoming increasingly popular among a variety of users, including professional translators and readers interested in obtaining the gist of texts in a foreign language. While any type of user can benefit from having high quality translations, this issue is much more crucial for a specific and very common type of user: those who cannot understand the language of the source text. These constitute a large proportion of users of MT systems, particularly online translation systems. These users are generally able to identify problems that affect the fluency of the translations, but not the less evident and more significant problems, such as the incorrect translation of an ambiguous word, the incorrect assignment of semantic roles in a sentence, or a reference to an incorrect antecedent. For example, consider the example in Figure 1, which contains a translation by an SMT system, followed by the source text and a human translation.

Target:
the road boycotted a friend ... indian robin hood killed the poor after 32 years of prosecution.

Source:
مقتل روبن هود الهندي.. قاطع الطريق صديق الفقراء بعد 32 عاما من الملاحقة

Reference:
death of the indian robin hood, highway robber and friend of the poor, after 32 years on the run.

Figure 1: Example of English MT for an Arabic source sentence and its reference translation

Fluent but inadequate translations such as the one in the example are commonly produced by SMT systems, given the usually strong bias of the language model component towards choosing a translation that is common (and thus fluent) in the target language, particularly in the absence of enough statistics for the translation model component.

We propose an approach to inform the end-users about the **adequacy** of a translation for a given input segment (sentence), so that the user is able to

judge whether or not to rely on the information in that translation. Such a mechanism to inform end-users who are not able to identify adequacy issues in the translation is crucial to avoid information misinterpretation.

Different from previous work, the approach proposed here is based on human assessments for adequacy and a number of translation quality indicators to contrast the source and translation texts. These range from simple frequency information about tokens in the source and target sentences to different levels of linguistic information. Experiments with Arabic-English translations show that the proposed prediction models can yield more reliable adequacy estimators for new translations.

In Section 2 we present related work in the field of quality estimation for MT. In Section 3 we describe the proposed approach, the datasets, features, resources and algorithms used. In Section 4 we present our experiments and results.

2 Related Work

Most work on sentence-level quality estimation (QE) – also called *confidence estimation* – proposed so far has focused on (i) estimating general quality scores - such as automatic metrics like BLEU (Papineni et al., 2002) - for tasks like n-best list reordering or MT system selection (Blatz et al., 2003; Quirk, 2004; Specia et al., 2009; Specia et al., 2010) and (ii) estimating post-editing effort (He et al., 2010; Specia and Farzindar, 2010; Specia, 2011).

The first significant effort towards sentence level quality estimation is presented in (Blatz et al., 2003). A large number of source, target and MT system features are used to train machine learning algorithms to estimate automatic metrics such as NIST (Doddington, 2002), which are then thresholded into binary scores to distinguish "good" from "bad" translations. The results were not very encouraging, possibly due to the fact that the automatic metrics used do not correlate well with human judgments at the sentence-level. In fact, Quirk (2004) showed that using a small set of translations manually labeled for quality it is possible to obtain models that outperform those trained on a larger set of automatically labeled translations.

Specia et al. (2009) use similar features to train a regression algorithm on larger datasets annotated by humans for post-editing effort. Satisfactory results were achieved when using the estimated scores for practical applications such as the selection of the best translation among alternatives from different MT systems (Specia et al., 2010).

He et al. (2010) propose using QE to recommend a translation from either an MT or a Translation Memory (TM) system for each source segment for post-editing. The QE model is trained on automatic annotation for Translation Edit Rate (TER) (Snover et al., 2006) and the goal is to predict the translation that would yield the minimum edit distance to a reference translation. Specia and Farzindar (2010) use TER to estimate the distance between machine translations and their post-edited versions (HTER). The estimated scores showed to correlate very well with human post-editing effort. Subsequently, Specia (2011) focuses on a more objective type of annotation: post-editing time. This has shown to be the most useful to allow ranking translations according to the post-editing effort they require.

A recent direction in QE is the addition of linguistic information as features. Focusing on word-error detection through the estimation of WER, Xiong et al. (2010) use POS tags of neighbor words and a link grammar parser to indicate words that are not connected to the rest of the sentence. Bach et al. (2011) check whether the dependency relations in the source sentence are preserved in the translation. Both approaches have shown the potential of linguistic features, but only Bach et al. (2011) use features contrasting the source and translation texts. However, these papers either focus on word-level quality estimation or on the estimation of automatic evaluation metrics. Moreover, they do not distinguish the types of errors in terms of fluency and adequacy: a substitution error referring to a simple morphological variation (with no effect on adequacy) is considered in the same way as a content word substitution changing the meaning of the sentence.

Framing the problem of QE as an adequacy estimation problem requires two main components: (i) new features that can better reflect aspects that have an impact on adequacy, and (ii) appropriate labeling of translation examples in terms of adequacy to train machine learning algorithms.

3 Adequacy Estimation Approach

We follow the standard approach for quality estimation: a machine learning algorithm trained on previously assessed translations and a number of quality indicators. However, we focus on adequacy indicators and explicit human annotations for adequacy.

An "adequate" translation can be defined as a translation that preserves the meaning of the input text and does not add any information to it. A fluent translation, on the other hand, is a grammatical and natural text in the target language. Adequacy and fluency are generally the two most desirable features for a correct translation. While quantifying these two aspects separately may not be straightforward, this has been the strategy used in some of the most relevant MT evaluation campaigns (Callison-Burch et al., 2010). For the application targeted in this paper, adequacy is more relevant than fluency. Disfluent translations can be identified by the reader without referring to the source text. Adequacy, on the other hand, can only be evaluated with respect to the source text, which makes it impossible for readers who cannot understand the source language.

While some of the features commonly used for QE can correlate reasonably well with adequacy, we believe more advanced features directly contrasting source and translation texts are necessary. Therefore, we identify a number of such features that can reflect (the lack of) adequacy to complement standard QE features. To distinguish these new features from existing ones, in Figure 2 we categorize different types of indicators used in our experiments.

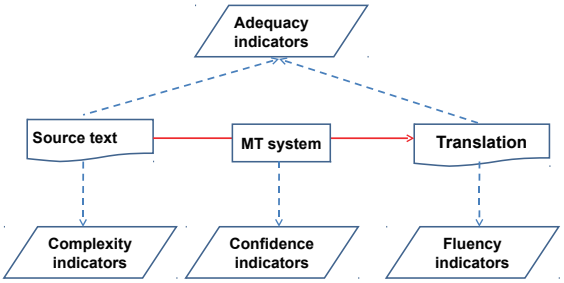

Figure 2: Categories of features for quality estimation

The vast majority of previous work has focused on (i) "confidence" indicators, i.e., features reflecting how confident the MT system is about the produced translation, such as the internal features of the SMT systems (phrase probability, distortion count, etc.), (ii) fluency indicators, i.e., features reflecting how natural and grammatical the translation is, and sometimes (iii) complexity indicators, i.e., features reflecting how difficult it is to translate the source text. In this paper, we focus on MT system-independent features, which can be extracted even if the user has no access to actual the MT system. This is a common scenario, particularly with online MT systems. In addition to complexity and fluency features, we propose a number of **adequacy** indicators, i.e., features that reflect how close or related the source and translation sentences are at different linguistic levels. In Section 3.2 we give examples of features in each of these categories.

3.1 Datasets and Adequacy Annotation

Three Arabic newswire datasets produced as part of the DARPA GALE project are used in this paper. Two state of the art phrase-based SMT systems, S1 and S2, were used to produce English translation for the datasets. Both systems were trained on a large parallel corpus of newswire texts ($\sim$ 6 million sentence pairs). Table 1 shows some statistics about these datasets.

Dataset	# Snt	# Words	METEOR	
			S1	S2
MT08	813	19,925	0.566	0.550
GALE09-dev	683	17,296	0.578	0.564
GALE10-dev	1,089	31,312	0.602	0.588

Table 1: Arabic-English datasets: number of sentences, (Arabic) words and corpus-level METEOR scores for SMT systems S1 and S2 using a single reference translation

In order to collect human annotations for translation adequacy, translations were given along with the source sentences to two Arabic-English professional translators. Each translation was annotated once (for each translation, one translator was randomly selected). The annotation was performed at a reasonably low cost: one US dollar per sentence.

Translators were asked to assess adequacy using a four point scale to answer the question: "To which degree does the translation convey the meaning of the original text?", where[1]:

[1]The complete guidelines include more details and examples

4 = Highly Adequate: The translation faithfully conveys the content of the input sentence. The translated sentence expresses exactly what the input sentence means. It reads perfectly to a speaker of English, or it needs some small corrections, but these could be done without referring to the Arabic input sentence. For example, there may be a problem with voice or number/tense/genre agreement, but it is clear from the English sentence what modifications are needed to correct such a problem without reading the Arabic sentence.

3 = Fairly Adequate: While the translation generally conveys the meaning of the input sentence, there are some problems with word order or tense/voice/number, or there are repeated, added or untranslated words. These problems partially change the meaning of the sentence. A speaker of English would be able to get the gist of the sentence, but some information would be missing or incorrect without referring to the Arabic input sentence.

2 = Poorly Adequate: The content of the input sentence is not adequately conveyed by the translation. There are problems with the relationships between words, clauses, or missing phrases or words, or with the polarity, incorrect translation of words or phrases, or other problems that significantly change the meaning of the sentence. A speaker of English would be able to get some information from the sentence, but the main message would be missing or incorrect.

1 = Completely Inadequate: The content of the input sentence is not conveyed at all by the translation. The meaning of the translation is different from that of the Arabic sentence, misleading the reader to a different interpretation, or the quality of the translation is so low that it is not a proper sentence and cannot be read.

The distribution of the scores for each of the MT systems and datasets as given by the translators is shown in Table 2. Since reference translations are also available as part of the GALE datasets, automatic metrics used in previous work can be considered as an alternative way of annotating the translations for quality. We chose METEOR (with lemmas, synonyms and paraphrases) (Denkowski and Lavie, 2010), as it has been shown to correlate better with the human perception of translation quality in previous work. Each sentence was annotated with its METEOR score computed using the reference translation. The average scores for each dataset are shown in Table 1.

Dataset	MT	**1** (%)	**2** (%)	**3** (%)	**4** (%)
MT08	S1	2.1	15.5	37.6	44.8
	S2	2.3	19.7	39.7	38.2
GALE09-dev	S1	2.3	22.8	47.4	27.4
	S2	3.2	22.0	53.9	20.9
GALE10-dev	S1	1.7	20.2	48.4	29.7
	S2	1.8	25.2	50.6	22.4

Table 2: Distribution of scores given by translators to each dataset and MT system

For feature extraction, all the datasets were preprocessed as follows:

Arabic (source): word transliteration, segmentation and morphological analysis using MADA (Habash and Rambow, 2005); POS tagging and chunking using AMIRA (Diab, 2009), constituent and dependency parsing using the Stanford parser (Green and Manning, 2010), and NER using a model learned from projections of English named entities (Section 3.2.1).

English (target): chunking using OpenNLP[2], constituent and dependency parsing using the Stanford parser (de Marneffe and Manning, 2008), NER using a combination of the Stanford (Finkel and Manning, 2010) and OpenNLP NER systems.

3.2 Features

The feature set used in this paper includes features from all categories shown in Figure 2. In total, 122 MT system-independent features were extracted for both S1 ans S2 datasets. In what follows we describe the adequacy features proposed in this paper, as well

of translations scored using the scheme

[2]http://incubator.apache.org/opennlp/

as provide some examples of the non-adequacy related features - please refer to (Blatz et al., 2003) for a complete list of source complexity and fluency features.

SF - Source complexity features:

- source sentence length
- source sentence type/token ratio
- average source word length
- source sentence 3-gram language model probability obtained based on the source side of the parallel corpus used to build the translation model of the SMT system

TF - Target fluency features:

- target sentence 3-gram language model probability obtained based on a large in-domain corpus of the target language
- translation sentence length
- coherence of the target sentence as in (Burstein et al., 2010)

AF - Adequacy features:

- ratio of number of tokens in source and target and vice-versa
- absolute difference between number of tokens and source and target normalized by source length
- ratio of percentages of numbers, content- / non-content words in the source & target
- ratio of percentage of nouns/verbs/etc in the source and target
- absolute difference between number of superficial constructions in the source and target: brackets, numbers, punctuation symbols
- proportion of dependency relations with constituents aligned between source and target
- absolute difference between the depth of the syntactic trees of the source and target
- absolute difference between the number of PP/NP/VP/ADJP/ADVP/CONJP phrases in the source and target
- difference between the number of 'person'/'location'/'organization' entities in source and target sentences

Source	Barack Obama [PERSON]
Target	bArAk AwbAmA [X]
Counts	0.7368 0.6829

Figure 3: A rule and its probabilities: $p(tag|en) = n(en, tag)/n(en)$ and $p(en|ar) = n(en, ar)/n(ar)$

- percentage of incorrectly translated direct object personal or possessive pronouns
- proportion of matching chunk labels in the source and target

3.2.1 Projecting Arabic Named Entities

The preservation of Named Entities (NE) is one of the desirable characteristics of a correct translation. Some of the features described in the previous Section are based on matching the number and type of entities in the source and target sentences. Since no freely available wide-coverage Named Entity Recognizer (NER) for Arabic exists, we implemented a simple model based on the projection of English NE obtained using a large Arabic-English in-domain parallel corpus. The English side of the parallel corpus is first annotated for NEs (Person, Location and Organization). We use both the Stanford (Finkel and Manning, 2010) and the OpenNLP NER systems[3]. The English annotations are projected to the Arabic side using word-alignment information. We align the parallel corpus using GIZA++ (Och and Ney, 2003) in the both directions (ar-en and en-ar) to produce the symmetrized alignment using tools provided by the Moses toolkit[4].

We then collect entities and their types to compute the context-independent probability distribution $p(ar|tag)$. More specifically, the word alignment and the source annotation is used to extract synchronous productions in a similar way to the rule extraction in tree-based SMT. The collection of annotated phrases is stored in a rule table with some relevant scores as exemplified by Figure 3.

We use the resulting rule table to estimate the probability of assigning a given entity type to an Arabic n-gram $p(tag|ar)$. As we do not have direct evidence of annotated Arabic entities we use the English translations of an Arabic string

[3]The OpenNLP NER is used to complement Stanford. If any conflict exists, Stanford NE are preferred.

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

Entity	Precision	Recall	F1
Person	85.42	61.19	71.3
Location	78.04	82.92	80.4
Organization	72.09	51.1	59.81
Average	77.92	67.15	72.13

Table 3: ANER performance on a tuning set

and their tags as pivot indicators: $p(tag|ar) = \sum_{en} p(tag|en)p(en|ar)$.

Once the entity probabilities are estimated, the model is applied to annotate the sentences in our datasets by identifying all the n-grams known to be possible entities and tagging them with their most frequent type.

We produced gold-standard annotations for 200 Arabic sentences which allowed us to tune a minimum acceptable $p(tag|ar)$ probability and a maximum entity length. Table 3 reports the best performance on this tuning set. Although the system does not consider the context of the NE, it performs reasonably well, since it was trained using in-domain data.

3.3 Learning Algorithms

To learn a quality prediction model, we use a Support Vector Machines (SVM) classification and regression algorithms with radial basis function kernel from the LIBSVM package[5] with the parameters γ, ϵ and *cost* optimized using a grid search approach.

Traditionally, quality estimation has been addressed as a binary problem: distinguishing correct from incorrect translations (Blatz et al., 2003). While this may be appropriate for certain applications, such as human post-editing of machine translations, for adequacy indication purposes, we believe a more detailed prediction is more informative. Therefore, we also train multi-class classifiers using three variations of the scores by grouping (or not) the initial scores in different ways:

Clas. 1: 4-class classifier to predict each of the four adequacy classes.

Clas. 2: a binary classifier to distinguish "adequate" (scores 3 and 4) from "inadequate" (scores 1 and 2) translations.

[5] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Clas. 3: a binary classifier to distinguish "fully adequate" (score 4) from "partially adequate or inadequate" (scores 1, 2 and 3) translations.

Additionally, we trained a regression algorithm on the dataset annotated with METEOR scores. The regressor is evaluated in terms of Root Mean Squared Error (RMSE): the average deviation from the predicted score to the expected score.

4 Results

Since our datasets are relatively small, we put them together to obtain a large enough number of instances for the SVM training. The resulting dataset containing 2,585 instances per MT system was randomly split into 85% for training and the remaining for test. This process was repeated 3 times to generate different splits. Table 4 shows the average accuracy of the classifiers for translations produced by each MT system in two settings: with (SF,TF,AF) and without (SF,TF) the adequacy features.

Clas	MT	SF,TF	SF,TF,AF	MC
1	S1	51.94 ± 3.2	51.51 ± 2.1	45.31 ± 2.1
	S2	51.42 ± 2.9	54.26 ± 4.3	48.49 ± 2.9
2	S1	79.85 ± 2.2	80.06 ± 2.0	79.24 ± 2.3
	S2	75.80 ± 0.8	75.97 ± 1.2	75.71 ± 0.7
3	S1	69.51 ± 3.1	69.51 ± 3.4	67.29 ± 2.0
	S2	73.56 ± 3.0	75.45 ± 2.9	72.78 ± 2.2

Table 4: Accuracy of the classifiers with all except adequacy features (source and target features: SF,TF), compared against all features (source, target and adequacy features: SF,TF,AF). MC (majority class) assigns the most common class in the training set to all test cases

The two variations of adequacy estimation models outperform the majority class classifier (MC), particularly for the 4-class classifier (Clas 1). However, only in some of our settings the models using the adequacy features yield better accuracy as compared to the models without such features. Overall, the performance of the classifiers seems strongly biased towards the majority class. This may be due to three main reasons: (i) the datasets are too small to contain enough instances of feature values for the large number of features used, (ii) the features are not sufficient to reflect adequacy or are too sparse, or (iii) the annotation scheme used for the datasets makes the distinction between the classes too difficult.

Datasets with similar sizes have been reported to be sufficient for post-editing effort estimation (Specia, 2011). However, since many features have been added here, larger datasets may be necessary.

Sparsity is an issue particularly with the adequacy features, given that not all linguistic phenomena tested happen in all sentences. For example, a feature checking the matching of named entities in the source and translation sentences will only have a relevant value for sentences containing named entities. We have tried a feature selection technique that ranks features according to their individual discriminative power (classifiers with a single feature) and builds classifiers adding the *k*-best features at a time until no improvement is found. However, the results were not significantly better.

Finally, the annotation scheme seems to be particularly complex because of category *3* = *Fairly Adequate*. While the distinction between this and its two adjacent categories (*4* = *Highly Adequate* and *2* = *Poorly Adequate*) seems clear to human translators, learning a model to distinguish *3* from *4* and *2* from *3* seems to require more complex features. E.g.: the distinction between *3* and *2* refers to whether the inadequacy is due to problems with the main message of the sentence or some satellite message. The main message of the sentence is not captured by the current features. This issue will be addressed in future work using less fine-grained adequacy categories and possibly additional features.

MT	SF,TF	SF,TF,AF
S1	0.0985 ± 0.005	0.0988 ± 0.003
S2	0.0956 ± 0.008	0.0941 ± 0.006

Table 5: Root mean squared error of the regression algorithm with all except adequacy features (SF,TF), compared against all features (SF,TF,AF)

Results using the regression algorithm (Table 5) are more positive. When METEOR is used for annotation, the adequacy features do not seem to contribute significantly to the other types of features. This was expected, given that METEOR is generally unable to distinguish the different levels of adequacy targeted by our features. Since METEOR varies from 0 to 1 (see average scores in Table 1), an average RMSE of ~ 0.1 represents a deviation of $\sim 10\%$ from the expected score. This low deviation is also shown in Figure 4, which contrasts the predicted and expected METEOR scores using all features and one of the MT system (S1) and test splits (the plots for all but adequacy features, other systems and splits are very similar).

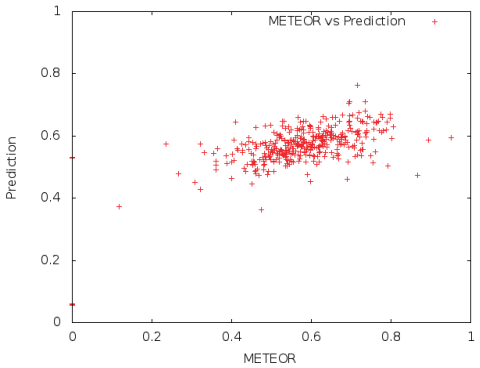

Figure 4: Expected versus predicted METEOR scores

5 Conclusions

We presented experiments with a number of novel translation adequacy indicators as part of a model to predict the adequacy of machine translations. The experiments demonstrate that estimating translation adequacy is a more complex problem than estimating the automatic metrics such as METEOR. The results achieved using adequacy annotations show consistent improvement with respect to a baseline (majority class), however, the contribution of the adequacy features is only evident in some of our testing conditions.

Further investigation is necessary to better understand the reasons for the relatively poor improvement in performance achieved with the adequacy indicators. This will include the following directions: (i) alternative ways of obtaining annotations for adequacy, and (ii) additional, possibly less-sparse adequacy indicators.

Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under GALE Contract No. HR0011-08-C-0110. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: A method for measuring machine translation confidence. In *49th Annual Meeting of the Association for Computational Linguistics*, pages 211–219, Portland, Oregon.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence estimation for machine translation. Technical report, Johns Hopkins University, Baltimore.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *North American Chapter of the Association for Computational Linguistics*, pages 681–684, Los Angeles, California.

Chris Callison-Burch, 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 *Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The stanford typed dependencies representation. In *Coling 2008 Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support for Five Target Languages. In *Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342, Uppsala, Sweden.

Mona T. Diab. 2009. Second generation amira tools for arabic processing: Fast and robust tokenization, pos tagging, and base phrase chunking. In *MEDAR 2nd International Conference on Arabic Language Resources and Tools*, pages 285–288, Cairo, Egypt.

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

Jenny R. Finkel and Christopher D. Manning. 2010. Hierarchical joint learning: improving joint parsing and named entity recognition with non-jointly labeled data. In *48th Annual Meeting of the Association for Computational Linguistics*, pages 720–728, Uppsala, Sweden.

Spence Green and Christopher D. Manning. 2010. Better arabic parsing: baselines, evaluations, and analysis. In *23rd Conference on Computational Linguistics*, pages 394–402, Beijing, China.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Annual Meeting on Association for Computational Linguistics*, pages 573–580, Ann Arbor, Michigan.

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

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

Kishore Papineni, 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, PA.

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

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *7th Conference of the Association for Machine Translation in the America*, pages 223–231, Cambridge, MA.

Lucia Specia and Atefeh Farzindar. 2010. Estimating Machine Translation Post-Editing Effort with HTER. In *AMTA-2010 Workshop Bringing MT to the User: MT Research and the Translation Industry*, pages 33–41, Denver, Colorado.

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

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24:39–50.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *15th Conference of the European Association for Machine Translation*, pages 73–80, Leuven, Belgium.

Deyi Xiong, Min Zhang, and Haizhou Li. 2010. Error detection for statistical machine translation using linguistic features. In *48th Annual Meeting of the Association for Computational Linguistics*, pages 604–611, Uppsala, Sweden.