Mach Translat (2010) 24:39–50
DOI 10.1007/s10590-010-9077-2

Machine translation evaluation versus quality estimation

Lucia Specia · Dhwaj Raj · Marco Turchi

Received: 15 May 2009 / Accepted: 21 April 2010 / Published online: 14 May 2010

Abstract Most evaluation metrics for machine translation (MT) require reference translations for each sentence in order to produce a score reflecting certain aspects of its quality. The de facto metrics, BLEU and NIST, are known to have good correlation with human evaluation at the corpus level, but this is not the case at the segment level. As an attempt to overcome these two limitations, we address the problem of evaluating the quality of MT as a prediction task, where reference-independent features are extracted from the input sentences and their translation, and a quality score is obtained based on models produced from training data. We show that this approach yields better correlation with human evaluation as compared to commonly used metrics, even with models trained on different MT systems, language-pairs and text domains.

Keywords Machine translation evaluation · Quality estimation · Confidence estimation

Lucia Specia—Work developed while working at the Xerox Research Centre Europe, France.
Dhwaj Raj—Work developed during an internship at the Xerox Research Centre Europe, France.
Marco Turchi—Work developed while working at the Department of Engineering Mathematics, University of Bristol, UK.

L. Specia (✉)
Research Group in Computational Linguistics, University of Wolverhampton, Wolverhampton, UK
e-mail: l.specia@wlv.ac.uk; lspecia@gmail.com

D. Raj
Indian Institute of Information Technology, Allahabad, India
e-mail: dhwaj@ug.iiita.ac.in

M. Turchi
European Commission – JRC (IPSC), T.P. 267, 21020 Ispra, Italy
e-mail: marco.turchi@bristol.ac.uk

1 Introduction

Recent shared evaluation tasks have shown progress on the average quality of machine translation (MT) systems, particularly in the case of statistical approaches (SMT) (Callison-Burch et al. 2009). The quality is measured based on both human evaluation and automatic metrics, which compare system translations against reference translations. While the most commonly used metrics like BLEU (Papineni et al. 2002), NIST (Doddington 2002) and Meteor (Lavie and Agarwal 2007) measure the overlap of n-grams between system and reference translations, recently proposed metrics like ULC (Gimenez and Marquez 2008) go beyond the lexical level to consider also the overlap of other linguistic features, like syntactic constituents, semantic roles, and discursive representations.

Provided that reference translations are available, these metrics can be used to evaluate the output of any number of systems, without the need for human intervention. The reliability of such metrics is usually measured as their correlation with human judgments, since the ultimate goal of any metric in this setting is to approximate human judgments accurately. While the de facto metrics—BLEU and NIST—correlate well with human judgments at the system level, new metrics like ULC can be optimized for sentence level evaluation. However, the need for reference translations certainly limits the amount of data that can be evaluated by using any of these metrics.

A problem that has been seen as orthogonal to MT evaluation is that of *predicting* the quality of machine translations. Sometimes referred to as confidence or quality estimation (QE), this task consists in estimating the quality of a system's output for a given input, without any information about the expected output, that is, without reference translations. Having an automatic way to assess the quality of translations is crucial from a practical point of view, when considering the interaction between an MT system and an end user (e.g. a professional translator). Without such an estimate, based on the system output only, the user can judge its fluency and maybe recognize clearly bad translations, but it is otherwise necessary to read the source text and the translation output to assess its quality. This is a very time consuming task and may not even be possible, if the user does not know the source language.

Traditionally, QE for MT has been viewed as a binary classification problem (Blatz et al. 2003) to distinguish between "good" and "bad" translations (of words, phrases or sentences). However, it may be difficult to find a clear boundary between "good" and "bad" translations and this information may not be useful for certain applications, for example, if one wants to estimate the effort necessary to post-edit translations.

In this paper we show that QE can be extended to predict not only a binary score but also a discrete or continuous score in a given range, and that this can be used in different applications. We investigate the problem of predicting the quality of translations at the sentence level when references are not available, using classification/regression algorithms and a large number of language-, resource- and MT-system independent features. Experiments with this method for translations produced by various MT systems and different language pairs yield estimates that correlate better with human judgments than metrics like BLEU and NIST. Besides contrasting QE with standard MT evaluation, we show that quality estimates can also be used for practical applications like filtering out bad translations for human post-editing and selecting the best translation from multiple MT systems.

In the remainder of this paper we first introduce related work on QE and MT evaluation (Sect. 2), then describe our experimental setting (Sect. 3) and present the results obtained (Sect. 4).

2 Related work

Early work on QE for MT aimed at estimating the quality at the word or phrase level (Gandrabur and Foster 2003; Ueffing and Ney 2005; Kadri and Nie 2006). The first comprehensive investigation on QE 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. However, there is no clear reason to believe that exactly the top 5 or 30% of translations are good and the remaining, bad.

Quirk (2004) uses classifiers and a pre-defined threshold for "bad" and "good" translations considering a small set of translations manually labeled for quality (350 sentences). Models trained on this dataset outperform those trained on a larger set of automatically labeled data.

Gamon et al. (2005) train a classifier using linguistic features extracted from machine and human translations to distinguish between these two types of translations (*human-likeness classification*). The predictions obtained have very low correlation with human judgments, which is an indication, as shown in Albrecht and Hwa (2007a), that high human-likeness does not necessarily imply good MT quality and vice-versa.

Some of the recently proposed metrics for sentence-level MT evaluation also exploit learning techniques and/or try to avoid the use of reference translations. For example, Albrecht and Hwa (2007b) use a regression algorithm with features extracted from MT output and *pseudo-references*. Pseudo-references are translations produced by other MT systems, instead of human references, but this is still essentially a reference-based evaluation scenario, where alternative MT systems are necessary. Pado et al. (2009) use a regression algorithm and features based on textual entailment between the translation and the reference. All these metrics are, therefore, based on some kind of reference-dependent features.

3 Experimental setup

3.1 Features

A number of features have been used in previous work for QE (see Blatz et al. (2003) for a list). In order to perform the task of QE across different MT systems, which may use different frameworks, we focus on features that do not depend on any aspect of the translation process, that is, which can be extracted from any MT system, given only the input (source) and MT output (target) sentences, and possibly (external) monolingual or parallel corpora. We call these "black-box" features.

We use most of the MT system-independent features that have been proposed in previous work and also identify a number of new features. In what follows, we describe the set of 74 features used in this paper, grouped here for space reasons. New features with respect to previous work are signalized by '*'.

- source & target sentence lengths and their ratios
- source & target sentence 3-gram language model probabilities and perplexities
- source & target sentence type/token ratio
- average source word length
- percentage of 1 to 3-grams in the source sentence belonging to each frequency quartile of a monolingual corpus
- number of mismatching opening/closing brackets and quotation marks in the target sentence
- average number of occurrences of all target words within the target sentence
- alignment score (IBM-4) for source and target sentences and percentage of different types of word alignments, as given by GIZA++ using the actual SMT training data ($\sim$1 million sentences) plus the QE sentences (*)
- average number of translations per source word in the sentence (as given by probabilistic dictionaries), unweighted or weighted by the (inverse) frequency of the words (*)
- percentages of numbers, content-/non-content words in the source & target sentences (*)
- percentages and number of mismatches of each of the following superficial constructions between the source and target sentences: brackets, punctuation symbols, numbers (*)
- 3-gram target language model probability trained on a corpus of POS-tags of words (*)

3.2 Data

We use translation data produced by four MT systems: Matrax (Simard et al. 2005), Portage (Johnson et al. 2006), Sinuhe (Kääriäinen 2009) and maximum margin regression (MMR) (Saunders 2008). Portage and Matrax are standard phrase-based SMT systems, with the exception that Matrax allows for gaps in phrases. Sinuhe is also a phrase-based system, but differs from standard systems by allowing phrases to overlap during decoding, and by training individual phrase weights applying a regularized Conditional Random Field on the full parallel aligned corpus. MMR is a rather distinct approach to MT based on using predictions with structured output. It is an end-to-end translation system that does not rely on traditional alignment or language-modeling tools. Features based on global similarities of words are first calculated using a minimum-distance approach on sentence pairs. Then, a structured-learning approach is used to compute the alignment of words to phrases. Decoding is performed by dynamic programming and is guided by a heuristic based on overlapping bigrams. Sinuhe and MMR were still at the initial development stages when used to produce the translations. In the following sections we anonymize these systems by arbitrarily naming them System-1 to System-4.

Two types of datasets are produced by these four systems:

1. en–es datasets: each system is trained on approximately 1 million English–Spanish sentence pairs from the Europarl corpus as provided by WMT-08 (Callison-Burch et al. 2008) and used to translate 4K Europarl sentences from the development and test sets also provided by WMT-08.
2. en–dk dataset: one MT system is trained on approximately 200,000 English–Danish sentence pairs from a technical corpus on the automotive industry domain and used to translate 3K sentences of the same domain. These sentences are very different from those in Europarl: they are shorter (10 words on average) and use simpler vocabulary and syntax (usually direct instructions on how to assemble, fix, etc., car components).

Translations produced by each system were manually annotated by professional translators with 1–4 quality scores, which is a range commonly used by them to indicate the quality of translations with respect to the need for post-editing:

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

Since different translators would be needed to annotate the datasets, we performed a pilot annotation task in order to verify the agreement among them. The same 50 en–es translations produced by one of the systems were given to three translators. The $Kappa$ (Cohen 1960) score obtained was 0.65, which is considered substantial. The datasets and their average human score are shown in Table 1.

The en–es datasets, that is, four datasets of 4,000 {source, translation, reference, human-score} quadruples produced by the four English–Spanish SMT systems can be downloaded from http://pers-www.wlv.ac.uk/~in1316/resources/ce_dataset.rar. They can be used for assessing existing MT evaluation metrics and also investigating new metrics based on human evaluation.

The feature vector for each dataset is randomly subsampled in training (75%) and test (25%) using a uniform distribution. Identical subsamples are created for the en–es datasets, in order to investigate the applicability of our QE metric for selecting the best

Table 1 Datasets used in the experiments, with the average human score assigned to the sentences in [1, 4]

Language-pair	MT system	No. sentences	Average score
en–es	System-1	4,000	2.835
en–es	System-2	4,000	2.558
en–es	System-3	4,000	2.508
en–es	System-4	4,000	1.338
en–dk	System-3	3,000	2.745

translation of a given sentence from multiple MT systems. Optimization of parameters of the machine learning algorithms is performed by cross-validation using five different subsamples of the training set.

3.3 Regression and classification algorithms

Different implementations of support vector machines (SVM) were used in our experiments:

1. 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.
2. Binary classification: C-support algorithm with radial basis function kernel from the LIBSVM package with the γ and *cost* parameters optimized.
3. Multi-class classification: SVM^{Struct} with radial basis function from the SVM^{Light} package (Joachims 1999), with the parameters γ and *cost* optimized.

4 Results

Below we describe the outcome of different experiments with our approach, divided into three groups, according to the application of the predicted score.

4.1 QE for MT evaluation

In order to show how our QE metric performs as compared to MT evaluation metrics, we produce models using the SVM regressor and classifiers from each of our five training sets. We then apply the models on each of the five test sets, including those containing translations produced by different MT systems, and also for texts of different domain and language-pair. Finally, we compute the correlation coefficients of the QE scores with the human annotations ([1–4] scores), and compare them to the correlation of MT evaluation metrics with the human annotations.

In Table 2 we contrast the (absolute) Pearson's correlation of the QE score obtained by SVM regression against four commonly used MT evaluation metrics with a single reference for each test set (900 cases for en–es and 700 cases for en–dk): BLEU, NIST, TER (Snover et al. 2006) and Meteor. We use a smoothed version of BLEU with bi-grams only (Lin and Och 2004), since the usual implementation with 4-grams resulted in many zero-scored sentences and therefore a lower overall correlation score. The version of Meteor considers the stems of words instead of their surface form, which is known to correlate better with human evaluation.

Table 2 shows that the correlation of the score predicted using our method (last column) is superior to that of any MT evaluation metric. The differences are statistically significant with 99.8% confidence, according to bootstrapping re-sampling and paired t-test (Koehn 2004). The QE score also outperforms the other metrics with models trained on translations produced by a different MT system for the same type of data (Europarl en–es). More important, this is also holds for translations produced by the

Table 2 Pearson's correlation of automatic metrics and QE score obtained by SVM regression with human annotation

Test set	BLEU-2	NIST	TER	Meteor	Training set	QE score
en–es System-1	0.237	0.203	0.194	0.277	en–es System-1	**0.556**
					en–es System-2	**0.569**
					en–es System-3	**0.545**
					en–es System-4	**0.489**
					en–dk System-3	**0.391**
en–es System-2	0.209	0.195	0.168	0.240	en–es System-1	**0.531**
					en–es System-2	**0.562**
					en–es System-3	**0.547**
					en–es System-4	**0.442**
					en–dk System-3	**0.400**
en–es System-3	0.296	0.254	0.268	0.337	en–es System-1	**0.478**
					en–es System-2	**0.517**
					en–es System-3	**0.542**
					en–es System-4	**0.423**
					en–dk System-3	**0.390**
en–es System-4	0.165	0.129	0.145	0.231	en–es System-1	**0.481**
					en–es System-2	**0.494**
					en–es System-3	**0.482**
					en–es System-4	**0.524**
					en–dk System-3	**0.444**

same or different MT systems on a totally different type of data (technical documents) and different language-pair (en–dk). Figures for Spearman's correlation, which considers only the ranking of the scores, are higher for all datasets, but the proportion of the differences between the QE score and other metrics is similar to that obtained for Pearson's correlation.

The results for multi-class classification were less encouraging. For comparison, we show in Table 3 the Pearson's correlation obtained for each dataset (models trained on the same system data only, for space reasons, and same splits for training and test as in Table 2). The results with the regression method (Table 2) are consistently superior to these, even when different models are used. One reason might be the unbalanced distributions of scores in some datasets, particularly in the case of System-4 (the true score is 1 for ∼73% of the cases). This shows that regression approaches seem to be more appropriate for this task. The correlation of traditional evaluation metrics with respect to human scores is the same as in Table 2, and is repeated here to show that using a classification algorithm sometimes results in lower correlation with human scores as compared to such metrics.

It is worth mentioning that in these experiments we used all 74 features, out of which 39 are hypothesis-independent, that is, are based on the source sentence only. Intuitively, these 39 features are aimed at measuring the difficulty of the source sentence. However, the difficulty of the source sentence has a strong impact on the quality

Table 3 Pearson's correlation between the QE score obtained by SVM multi-class classification and human annotation

MT system	BLEU-2	NIST	TER	Meteor	QE score
en–es System-1	0.237	0.203	0.194	**0.277**	0.170
en–es System-2	0.209	0.195	0.168	0.240	**0.359**
en–es System-3	0.296	0.254	0.268	0.337	**0.396**
en–es System-4	0.165	0.129	0.145	**0.231**	0.190

of the output translation. Therefore, the hypothesis-independent features have shown to be very important for the performance of the QE metric. Experiments where only the remaining 35 features are used resulted in a considerable degradation of the QE results shown in Tables 2 and 3.

Although the goal of the QE metric is not essentially the same as that of general reference-based MT evaluation metrics, the experiments described in this section show that it is possible to use a metric which has no access to reference translations and obtain a good correlation with human assessment. The QE metric is not meant to replace evaluation metrics, but instead to provide a way to assess quality when reference translations are not available. Moreover, the experiments show that the difficulty of the source sentence has a strong influence on the overall quality of the translations. This could be an indication that MT metrics should use more information regarding the input text in order to evaluate the output translations.

4.2 QE for selecting the best translation from multiple MT systems

While most MT systems produce a score for each output translation, this score is not an absolute quantification of the quality of the translation. Furthermore, perhaps more importantly, it is not comparable across different MT systems, since these might be based on different translation models.

Another potential application is therefore to use the QE score, produced independently for each MT system dataset, to select the best translation for a given source sentence from multiple alternatives produced by different MT systems. This can be thought of as a simplified version of the "system combination" task.

We experimented with this application and our four en–es datasets, where the same 900 source sentences had been translated by four MT systems. For each source sentence, we selected the "best" translation out of the four options as the one with highest QE score. Using this approach, an accuracy of 77% was obtained, that is, 77% of the sentences with the highest QE score also had the highest score according to human annotation.

This accuracy can be compared to that of always choosing the translations coming from the best system overall. As shown by Callison-Burch et al. (2009), using the overall quality of the MT system (human or BLEU system-level score, for example) is a good predictor of the sentence-level quality of translations, and in many cases correlates considerably better to human evaluation than sentence-level evaluation metrics. In our datasets, if we assume that translations produced by System-1 (the best system

Table 4 MT evaluation metrics for individual systems and results of system combination by using QE

MT system	BLEU	NIST	TER	Meteor
en–es System-1	0.3712	8.1052	47.674	0.3785
en–es System-2	0.3476	7.7315	50.047	0.3534
en–es System-3	0.3169	7.7337	50.037	0.3364
en–es System-4	0.1922	6.1725	63.281	0.2546
en–es system-combination	**0.3822**	**8.2068**	**46.599**	**0.3898**

on average, as shown in Table 1) are always the best ones, the accuracy drops to 54%. That shows that the QE score can play an important role in the system combination task.

Another way to show the potential of the QE score for system combination is by computing traditional system-level MT evaluation metrics for the translations chosen according to the QE score as compared to the translations of each system individually. Table 4 shows the system-level score obtained using the MT evaluation metrics for each en–es dataset individually, and after system combination taking the best translation according to the QE score. The scores for the translations resulting from system combination are significantly higher than those for any individual dataset (or lower, in the case of TER, an error rate metric), including the best system overall, *System-1* (98% confidence according to bootstrapping re-sampling test (Koehn 2004)).

We have also experimented with using only the 35 hypothesis-dependent features for this task, since the 39 source-dependent features are the same in all four systems. This resulted in 64% accuracy, which is considerably lower than the results with all features (77%). Therefore, the interaction of source-dependent features with the remaining features seems to have a strong impact on the quality of the QE score.

4.3 QE for filtering out bad translations

In this section we consider the traditional scenario envisaged for QE: a binary task where translations are classified as "good" or "bad". This could be directly applied by language-service providers (LSPs) using MT to select for post-editing only translations considered "good", and therefore prevent professional translators from spending time reading bad-quality translations before rejecting them.

A heuristic that is sometimes used by LSPs in order to decide whether or not to post-edit the output of an MT system is the length of the sentence: long segments usually result in lower quality translations as compared to shorter ones. We propose the use of the QE score instead of sentence length for this purpose. This is done by grouping the $\{1–4\}$ human scores into two classes and directly estimating "good"/"bad" indicators. For training purposes, scores $\{1, 2\}$ are considered "bad", while $\{3, 4\}$ are taken as "good".

QE models are trained and tested using the same type of data (MT system, language-pair and domain), since this is a more natural scenario for this application. We measure the performance of this task by classification accuracy, that is, the proportion of correctly classified test cases. We compare this accuracy to that of a

Table 5 Classification accuracy of the QE score obtained by SVM binary classification (scores $\{1,2\}$ are considered “bad”, while $\{3,4\}$ are considered “good”) compared to the baseline of the majority-class classifier, SVM regression, and the use of sentence length as filtering criterion

Dataset	SVM binary	SVM regression	Majority class	Snt. length
en–es System-3	**0.698**	0.529	0.519	0.356
en–es System-1	**0.768**	0.578	0.738	0.212
en–es System-2	**0.660**	0.521	0.565	0.294
en–es System-4	**0.935**	0.700	**0.935**	0.700
en–dk System-3	0.706	**0.723**	0.621	0.687

trivial classifier that always chooses the majority class in the training set. We also investigate estimating a continuous score in $[1,4]$ and then thresholding it according to the threshold that would correspond to the $\{1,2\}$ and $\{3,4\}$ groups in the training set. Finally, we also show the accuracy obtained using sentence length as classification criterion and a threshold of 12 words. This threshold in the sentence length was identified by the LSP that has performed the human translation assessments in this paper (namely Xerox Global Knowledge and Language Services) as the most adequate for filtering out presumably bad machine translations. Results are shown in Table 5.

Apart from *en–es System-4*, which had a very skewed distribution (93.5% of the cases belonging to the same class), the accuracy of the binary classifiers was significantly superior to the baseline of the majority class and the sentence length criterion. Directly estimating a binary score performs significantly better than estimating a [1,4] score and then thresholding it in two classes for all en–es datasets, while the opposite was observed for the en–dk dataset.

5 Discussion and conclusions

We presented an approach to estimate the quality of machine translation at the sentence-level, contrasted it to traditional MT evaluation metrics and shown that it can be used for different applications: filtering out bad translations for post-editing of machine translation and selecting the best translation from multiple MT systems.

From the MT evaluation perspective, results show that, given a model trained on data from any MT system, language-pair and text domain, it is possible to obtain quality estimates for any number of new sentences, since reference translations are not necessary, and these estimates correlate significantly better with human evaluation than reference-based metrics commonly used for MT evaluation.

The scores in $\{1,4\}$ used to annotate the quality of translations is not very different from the $\{1,5\}$ range used in standard shared evaluation tasks like the WMT workshops. However, in our case the scores correspond to an absolute quantification of the quality of the sentences, instead of a relative quantification, where translations from a number of systems are scored comparatively. Given that our goal is to estimate an absolute score, using datasets manually annotated in these shared tasks is likely to degrade the results. In fact, comparing our approach to previous work using such datasets, like Albrecht and Hwa (2007b) and Gimenez and Marquez (2008), is not straightforward, as they put together translations produced by different MT systems for a given language pair. This is not a problem for reference-based metrics, since features are extracted from comparisons of the target and reference sentences only, but it is crucial in our case, where a large number of features are based on the source sentence only, and would therefore be the same for multiple test cases. In Specia et al. (2009) we experimented with smaller datasets from WMT-2006, each for a given language-pair and MT system. Results were found to be lower than the ones presented here in terms of prediction accuracy, although they may still be acceptable for certain applications.

The type of translation quality that can be measured with predictive approaches like ours depends both on the criteria used for human annotation and the features used. The very simple features used in this paper performed well to measure the $[1, 4]$ scores in terms of post-editing needs. As future work we plan to investigate the use of more elaborate, language-dependent features, like those proposed by Gimenez and Marquez (2008) for this task. However, instead of contrasting between target and reference translations, we will investigate how they could be extracted to contrast source and target sentences.

We also plan to investigate the combination of the QE score with reference-based MT evaluation metrics like ULC (Gimenez and Marquez 2008), in order to improve such metrics. Alternatively, we could incorporate some of the well performing QE features into metrics based on learning methods to predict a quality score, such as Albrecht and Hwa (2007a).

References

Albrecht J, Hwa R (2007a) A re-examination of machine learning approaches for sentence-level MT evaluation. In: 45th meeting of the association for computational linguistics, Prague, pp 880–887

Albrecht J, Hwa R (2007b) Regression for sentence-level MT evaluation with pseudo references. In: 45th meeting of the association for computational linguistics, Prague, pp 296–303

Blatz J, Fitzgerald E, Foster G, Gandrabur S, Goutte C, Kulesza A, Sanchis A, Ueffing N (2003) Confidence estimation for machine translation. Technical report. Johns Hopkins University, Baltimore

Blatz J, Fitzgerald E, Foster G, Gandrabur S, Goutte C, Kulesza A, Sanchis A, Ueffing N (2004) Confidence estimation for machine translation. In: 20th coling, Geneva, pp 315–321

Callison-Burch C, Fordyce C, Koehn P, Monz C, Schroeder J (2008) Further meta-evaluation of machine translation. In: 3rd workshop on statistical machine translation, Columbus, pp 70–106

Callison-Burch C, Koehn P, Monz C, Schroeder J (2009) Findings of the 2009 workshop on statistical machine translation. In: 4th workshop on statistical machine translation, Athens, pp 1–28

Chang C, Lin C (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46

Doddington G (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Conference on human language technology, San Diego, pp 138–145

Gamon M, Aue A, Smets M (2005) Sentence-level MT evaluation without reference translations: beyond language modeling. In: 10th meeting of the European association for machine translation, Budapest

Gandrabur S, Foster G (2003) Confidence estimation for translation prediction. In: 7th conference on natural language learning, Edmonton, pp 95–102

Gimenez J, Marquez L (2008) A smorgasbord of features for automatic MT evaluation. In: 3rd workshop on statistical machine translation, Columbus, OH, pp 195–198

Joachims T (1999) Making large-scale SVM learning practical. In: Schoelkopf B, Burges CJC, Smola AJ (eds) Advances in Kernel methods—support vector learning. MIT Press, Cambridge

Johnson H, Sadat F, Foster G, Kuhn R, Simard M, Joanis E, Larkin S (2006) Portage: with smoothed phrase tables and segment choice models. In: Workshop on statistical machine translation, New York, pp 134–137

Kääriäinen M (2009) Sinuhe—statistical machine translation using a globally trained conditional exponential family translation model. In: Conference on empirical methods in natural language processing, Singapore, pp 1027–1036

Kadri Y, Nie JY (2006) Improving query translation with confidence estimation for cross language information retrieval. In: 15th ACM international conference on information and knowledge management, Arlington, pp 818–819

Koehn P (2004) Statistical significance tests for machine translation evaluation. In: Conference on empirical methods in natural language processing, Barcelona

Lavie A, Agarwal A (2007) METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: 2nd workshop on statistical machine translation, Prague, Czech Republic, pp 228–231

Lin CY, Och FJ (2004) ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In: Coling-2004, Geneva, pp 501–507

Pado S, Galley M, Jurafsky D, Manning CD (2009) Textual entailment features for machine translation evaluation. In: 4th workshop on statistical machine translation, Athens, pp 37–41

Papineni K, Roukos S, Ward T, Zhu W (2002) Bleu: a method for automatic evaluation of machine translation. In: 40th meeting of the association for computational linguistics, Morristown, pp 311–318

Quirk CB (2004) Training a sentence-level machine translation confidence measure. In: 4th language resources and evaluation conference, Lisbon, pp 825–828

Saunders C (2008) Application of Markov approaches to SMT. Technical report. SMART Project Deliverable 2.2

Simard M, Cancedda N, Cavestro B, Dymetman M, Gaussier E, Goutte C, Yamada K (2005) Translating with non-contiguous phrases. In: Conference on empirical methods in natural language processing, Vancouver, pp 755–762

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Conference of the 7th association for machine translation in the Americas, Cambridge, MA, pp 223–231

Specia L, Turchi M, Cancedda N, Dymetman M, Cristianini N (2009) Estimating the sentence-level quality of machine translation systems. In: 13th meeting of the European association for machine translation, Barcelona

Ueffing N, Ney H (2005) Application of word-level confidence measures in interactive statistical machine translation. In: 10th meeting of the European association for machine translation, Budapest, pp 262–270