Estimating the Quality of Translated User-Generated Content

Raphael Rubino[†‡], **Jennifer Foster**[†], **Rasoul Samad Zadeh Kaljahi**[†‡],
Johann Roturier[‡] and **Fred Hollowood**[‡]
[†]NCLT, School of Computing, Dublin City University, Ireland
{rrubino, jfoster, rkaljahi}@computing.dcu.ie
[‡]Symantec Research Labs, Dublin, Ireland
{johann_roturier, fhollowood}@symantec.com

Abstract

Previous research on quality estimation for machine translation has demonstrated the possibility of predicting the translation quality of well-formed data. We present a first study on estimating the translation quality of *user-generated content*. Our dataset contains English technical forum comments which were translated into French by three automatic systems. These translations were rated in terms of both comprehensibility and fidelity by human annotators. Our experiments show that tried-and-tested quality estimation features work well on this type of data but that extending this set can be beneficial. We also show that the performance of particular types of features depends on the type of system used to produce the translation.

1 Introduction

Quality Estimation (QE) involves judging the correctness of a system output given an input without any output reference. Substantial progress has been made on QE for Machine Translation (MT), but research has been mainly conducted on well-formed, edited text (Blatz et al., 2003; Ueffing et al., 2003; Raybaud et al., 2009; Specia et al., 2009). We turn our attention to estimating the quality of *user-generated content* (UGC) translation – a particularly relevant use of QE since the translation process is likely to be affected by the noisy nature of the input, particularly if the MT system is trained on well-formed text.

The source language content is collected from an IT Web forum in English and translated into French by three automatic systems. For each MT system, the produced translation is manually evaluated following two criteria: the translation comprehensibility and fidelity. We evaluate several feature sets on the UGC dataset including the baseline suggested by the organisers of the WMT 2012 QE for MT shared task (Callison-Burch et al., 2012) and a feature set designed to model typical characteristics of forum text.

The novel contributions of the paper are: 1) testing the WMT QE for MT Shared Task baseline feature set on the UGC dataset and demonstrating its portability, 2) introducing new features which contribute to significant performance gains on both QE tasks, and 3) building three different QE systems using three different MT systems and showing that the usefulness of a feature type depends on the MT system although better performance can be achieved by training a QE system on the combined output of the three systems.

The paper is organised as follows. Related work on QE for MT is described in Section 2, followed in Section 3 by a description of the dataset. We describe the QE features in Section 4 and present the results of our experiments in Section 5. A discussion of the results, as well as a comparison with previous work, are presented in Section 6. Finally, we conclude and suggest future work in Section 7.

2 Background

The main approach for QE in MT is based on estimating how correct MT output is through characteristic elements extracted from the source and the target texts and the MT system involved in the translation process. These elements, or features, are seen as predictive parameters that can be combined with machine learning methods to estimate binary, multi-class, or continuous scores. First applied at the word level (Gandrabur and Foster, 2003; Ueffing et al., 2003), QE for MT was then extended to the sentence level during a workshop in the same year (Blatz et al., 2003).

Many different feature sources have been used including surface features (segment length, punc-

International Joint Conference on Natural Language Processing, pages 1167–1173,
Nagoya, Japan, 14-18 October 2013.

tuation marks, etc.), language model features (perplexity, log-probability, etc.), word or phrase alignment features, n-best list features, internal MT system scores (Quirk, 2004; Ueffing and Ney, 2004), and linguistic features (Gamon et al., 2005; Specia and Gimenez, 2010). In a recent study, features based on the intra-language mutual information between words and backward language models were introduced (Raybaud et al., 2011). Other studies evaluate the gain brought by features extracted from MT output back-translation (Albrecht and Hwa, 2007), pseudo-references in the form of output from other MT systems for the same source sentence (Soricut et al., 2012), and topic models (Rubino et al., 2012).

Previous studies also differ on the labels to predict: binary scores (Quirk, 2004) or continuous scores such as those given by automatic metrics (Bojar et al., 2013) or averaged human evaluations (Specia et al., 2009; Callison-Burch et al., 2012). As regards the learning algorithms used, several have been tried, with support vector machine and decision tree learning proving popular (Callison-Burch et al., 2012).

3 Dataset

We use the dataset presented in Roturier and Bensadoun (2011), which was obtained by machine-translating 694 English segments, harvested from the Symantec English Norton forum[1], into French using three different translators (MOSES (Koehn et al., 2007), MICROSOFT[2] (MS) and SYSTRAN). The translations were then evaluated in terms of comprehensibility (1 to 5 scale) and fidelity (binary scale) by human annotators. The source side of this data set represents user-generated content – see Banerjee et al. (2012) for a detailed description of the characteristics of this type of data and see Table 2 for some examples. For each of the three translators, we extract 500 segments from this dataset to build our training sets. The remaining 194 segments per translator are used as test sets. The distribution of the comprehensibility and fidelity classes over the three MT systems are shown in Table 1.

4 Quality Estimation Features

In this section, we describe the features which we added to the 17 baseline features provided by the WMT12 QE shared task organisers to make our "extended" feature set. We then introduce a set of 37 features which relate specifically to the user-generated-content aspect of our data.

	Comprehensibility					*Fidelity*
Class	1	2	3	4	5	1
MOSES	6.1	55.0	12.7	11.2	15.0	37.2
MS	10.7	39.8	19.2	13.5	16.9	46.0
SYSTRAN	11.5	45.7	14.8	11.7	16.3	41.2

Table 1: Distribution (%) over the comprehensibility and fidelity classes for the 694 segments per MT system.

4.1 Extended Feature Set

- 15 **Surface Features** Average target word length, average source word occurrence, number of uppercased letters and the ratio of all source and target surface features.
- 180 **Language Model Features** Source and target n-gram ($n \in [1;5]$) log-probabilities and perplexities on two LMs built on the seventh version of Europarl and the eighth version of News-Commentary (30 features). The same number of features are extracted from a backward version of these two LMs (Duchateau et al., 2002). We repeat this feature extraction process using four LMs built on the Symantec Translation Memories (TMs)[3] and four LMs built on the monolingual Symantec forum data[4].
- 15 **MT Output Language Model Features** A MOSES English-French PB-SMT system is trained on the Symantec TMs and the same target LM used to extract the baseline LM features. The English side of the Symantec Norton monolingual forum data is translated by this system and the output is used to build a 5-gram LM. Target features are then extracted in a similar way as the standard LM features.
- 4 **Word Alignment Features** Using GIZA++ (Och and Ney, 2000) and the Symantec TMs, word alignment probabilities are extracted from the source and target segments.
- 78 **n-gram Frequency Features** The number of source and target segments unigrams seen in a reference corpus plus the percentage of n-grams in frequency quartiles ($n \in [1;5]$). The reference corpus is the same corpus used to extract the LM features.

[1] http://community.norton.com

[2] http://www.bing.com/translator/

[3] $\sim 1.6M$ aligned segments in English and French.

[4] $\sim 3M$ segments in English and $\sim 40k$ segments in French.

so loe and behold I get a Internet Worm Protection Signature File Version: 20090511.001. on 5/20/09 in the afternoon
Start NIS 2009 > In the Internet pane, click Settings > Under Smart Firewall, click configure next to Advanced settings > In the Advanced Settings window, turn off Automatic Printer Sharing control.
ok then what should do am i safe as is meaning just leave it alone as long it get blocked it can get my info right i have no clew how get rid of it the only thing i could do that i know would work is to take everthing out my computer and format it with boo disk will this work?

Table 2: Processing challenges associated with forum text: some examples.

- 9 **Back-translation Features** We translate the target segments back into the source language using 3 different MT systems: MS, Systran and a Moses PB-SMT system (trained on the Symantec TMs) and we measure the distance between the original source segments and the back-translated ones using BLEU, TER and Levenshtein.
- 23 **Topic Features** Following Rubino et al. (2012; 2013), a bilingual topic model based on Latent Dirichlet Allocation (Blei et al., 2003) is built using the Symantec TMs. 20 features are the source and target segment distributions over the 10-dimensional topic space and 3 features are the distances between these distributions, using the cosine, euclidean distance and city-block metrics.
- 16 **Pseudo-reference Features** Following Soricut et al. (2012), we compare each MT system output to the two others using sentence-level BLEU, error information provided by TER (no. of insertions, deletions, etc.) and Levenshtein.
- 3 **Part-of-Speech Features** We count the number of POS tag types in the source and target segments, extracted from trees produced by the Stanford parser (Klein and Manning, 2003). The ratio of these two values is also included.

4.2 UGC-related Features

We also experiment with features that capture the noisy nature of UGC. Some are related to the inconsistent use of character case, some to non-standard punctuation, some to spelling mistakes and some to the tendency of sentence splitters to underperform on this type of text. From each source-target pair, we extract the following information (in the form of one feature for the source segment, one for the target segment and, where appropriate, one for the ratio between the two):

- 11 **Case Features** the number of upper and lowercased words, the number of fully uppercased words, the number of mixed-case words and whether or not the segment begins with an uppercase letter.
- 13 **Punctuation Features** the ratio between punctuation characters and other characters, the number of words containing a full stop, the number of sentences produced by an off-the-shelf sentence splitter for each segment (included in NLTK (Bird, 2006)), whether or not the segment contains a dash, an ellipsis, and whether or not the segment ends with a punctuation symbol.
- 9 **Acronym and Emotion Features** the number of web and IT-domain acronyms and the number of emoticons.
- 4 **Linguistic Features** the number of spelling mistakes flagged by the spellchecker LanguageTool[5] and whether or not the segment starts with a verb (indicating imperatives or questions).

5 Experiments

Classification models are built using the C-SVC implementation in LibSVM (Chang and Lin, 2011) with a Radial Basis Function (RBF) kernel. Optimal hyper-parameters C and γ are found by grid-search with a 5-fold cross-validation on the training set (optimising for accuracy). For evaluation, we measure the accuracy, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). All the results are compared to the baseline for significance testing using bootstrap resampling. We present the results on the comprehensibility task first, followed by the results on the fidelity task. In order to remove noisy and redundant features we also experiment with feature selection. We try several approaches[6] and report results with the approach that performs best during cross-validation on the training set.

5.1 Translation Comprehensibility Results

The full set of comprehensibility estimation results are presented in Table 3. We see that a higher classification accuracy does not necessarily imply lower MAE and RMSE, e.g. the Moses and

[5] http://www.languagetool.org/

[6] These include information gain univariate filtering, correlation-based multivariate filtering , a naive Bayes wrapper approach and principal component analysis. All are implemented in the WEKA machine learning toolkit (Hall et al., 2009). Of the approaches tried, none stood out as clearly superior to the others and the choice seems to depend on the task and the MT system.

SYSTRAN experiments show that the extended set leads to a higher classification accuracy compared to the baseline, while the two error scores are lower on the baseline compared to the extended set. This discrepancy can happen because the accuracy measure is not sensitive to differences between comprehensibility scores whereas the error measures are – the error measures will prefer a system which gives a 5-scoring translation a score of 4 than one which gives it a score of 1. Statistical significance tests show that the extended feature set outperforms the baseline significantly only for the system trained on MS translations. The UGC feature set seems to add useful information only for the system trained on MOSES translations.

	MOSES	MS	SYSTRAN
Baseline			
Acc. (%)	67.0	39.7	55.2
MAE	0.48	0.96	0.64
RMSE	**0.94**	1.38	1.07
Extended			
Acc. (%)	69.1	42.8⋆	55.7
MAE	0.51	0.87	0.65
RMSE	1.01	1.28	1.11
Extended+UGC			
Acc. (%)	69.1	41.8	55.2
MAE	0.49	0.89	0.67
RMSE	0.99	1.29	1.14
Extended + Feature Selection			
Acc. (%)	**70.6**⋆	39.2	**56.7**
MAE	**0.47**	0.92	**0.56**⋆
RMSE	0.97	1.32	**0.96**⋆
Feature Types + Feature Selection			
Acc. (%)	69.6	42.8⋆	52.6
MAE	0.49	0.85⋆	0.70
RMSE	0.99	**1.24**⋆	1.14
Mixed-Translator: Extended+UGC			
Acc. (%)	69.1	**44.3**⋆	54.6
MAE	0.48	**0.84**⋆	0.65
RMSE	0.98	1.27	1.09

Table 3: Translation comprehensibility estimation. Best results are in bold, statistically significant improvements over the baseline ($p < 0.05$) are indicated with ⋆.

To evaluate the impact of different types of features, we conduct an evaluation by feature subset (see Figure 1). The results show that the best-performing features vary across the MT systems. The pseudo-reference and POS features are particularly useful for the system trained on MOSES translations. For the system trained on MS translations, the n-gram frequency features based on the Symantec TMs are clearly outperforming all the other feature types, with an accuracy of 42.8%. For the system trained on SYSTRAN translations, the pseudo-reference features yield an accuracy score of 56.7%, which outperforms the extended set. The system trained on MS translations does not benefit from the pseudo-reference features as much as the two other QE systems. Perhaps this is because these features provide a reliable indication of translation quality when the two MT systems being compared are trained on similar data – MOSES was trained using the domain- and genre-specific data and SYSTRAN is optimized using a domain-specific lexicon, which increases the proximity of the translations generated by these two systems. LM features appear to be very useful for the three MT systems: the backward LM built on the TMs leads to the best accuracy results amongst the LM-based features for MS, while SYSTRAN and MOSES benefit from features extracted using a backward LM built on forum data.

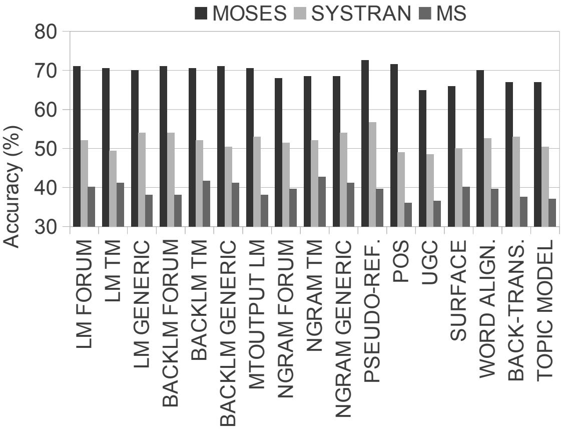

Figure 1: Feature types for comprehensibility.

According to the results in Fig. 1, several feature types individually outperform the baseline and the extended sets, which indicates that unsuited features are included in these two sets and motivates the application of feature selection. The feature selection algorithms are applied in two ways: on the extended set and on each feature type individually. For this second approach, the reduced feature types are combined to form the final set. The results obtained with the first and second feature selection methods are presented in the fourth and fifth rows of Table 3 respectively. The systems trained on MS and SYSTRAN translations clearly benefit from the feature selection process with significant improvement over the baseline. For the system trained on the MOSES translations, only the accuracy scores are improved over the baseline. The choice of which of the two methods of applying feature selection to use also depends on MT system.

As the training set for each MT system is small (500 instances), we combine these training sets and build a mixed-translator classification model (last row in Table 3). Note that this means that each training source segment will appear three times (one for each of the MT systems). We use the *Extended+UGC* feature set to build the mixed-translator model. Comparing to the results in the third row (individual MT classification model), we observe that it is generally beneficial to combine the translations into one larger model.

5.2 Translation Fidelity Results

The fidelity results are presented in Table 4.[7]

	MOSES	MS	SYSTRAN
Baseline	81.4	62.9	68.0
Extended	81.4	63.4	**76.8**
Extended+UGC	80.4	65.5	73.7
Extended+sel.	77.3	64.4	72.2
Type+sel.	**82.0**	**69.1**	74.2
Mixed-Translator: Ext+UGC	**82.0**	66.5	76.3

Table 4: Accuracy for fidelity estimation, best results are in bold.

According to the results for the extended feature set, the baseline result for the system trained on MOSES translations appears to be very difficult to improve upon. For this system, adding the UGC features actually degraded the accuracy scores, while it helps the system trained on MS translations. The extended set reaches the best accuracy scores (76.8%) for the system trained on SYSTRAN translations with a 8.8pt absolute improvement over the baseline set. However, statistical significance testing show that none of the improvements over the baseline are statistically significant.

As with the comprehensibility results, the impact of the feature sets depends on the MT system (see Figure 2). Again, the pseudo-reference features lead to the highest accuracy score for the system trained on MOSES translations (+2.1pts absolute compared to the baseline and extended sets) and the system trained on SYSTRAN translations (equal to the extended set), while it is not the case for the system trained on MS translations. For this latter system, the n-gram frequency features based on the forum data reach 66.0% accuracy. The accuracy results per feature type show a larger divergence compared to the results obtained on the comprehensibility task. Some feature types appear to be particularly noisy for the fidelity task, for instance the surface features for MOSES, the back-translation for MS and the word-alignment for SYSTRAN. To tackle this issue, the same feature selection methods previously used for the comprehensibility task are applied. The fourth and fifth rows in Table 4 show the results for the two methods of applying feature selection. We can see that selecting features within individual feature types leads to better results compared to applying feature selection to the full set.

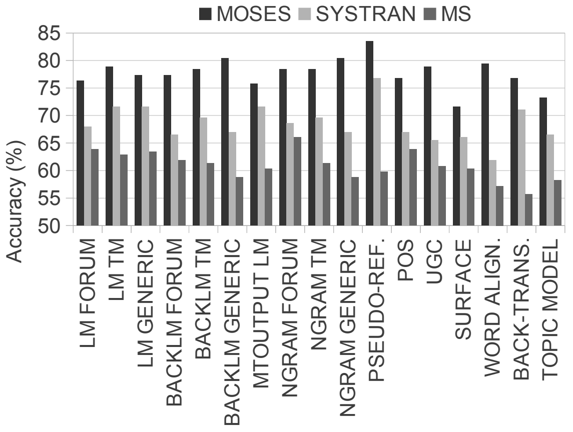

Figure 2: Feature types for fidelity.

As with the comprehensibility task, we build a mixed-translator fidelity estimator using the *Extended+UGC* feature set (last row in Table 4) and we observe here also that it is beneficial to combine the training data compared to training individual models.

6 Analysis

Adding features to the baseline set does not necessarily lead to better QE. Perhaps the baseline feature set is already diverse enough (surface, LM, word alignment, etc.). However, an error analysis shows that including the UGC features does bring useful information, especially when the source segments contain URLs, as shown in Table 5. In the case of untranslated elements, the spellchecker sometimes provides important information to the classifier about the MT output quality.

When we compare the QE results over the three MT systems, there is substantial variation. One possible explanation for this variation is the class distributions for the three sets of translations. The set whose quality is hardest to predict (MS) is the one with a more balanced distribution over the classes for both tasks (see Table 1). As classifiers

[7] We do not measure the two error scores for the fidelity scores prediction because it is a binary classification task.

Translator: MS	
Source	How to remove status bar indicator? <URL>
Target	Comment supprimer l'indicateur de la barre de statut ? <URL>
	Baseline → 2 **+UGC** → 4 **Ref** → 4
Source	Best Regards Anders
Target	Best Regards Anders
	Extended → 5 **+UGC** → 1 **Ref** → 1
Translator: SYSTRAN	
Source	If you look at the URL of a Norton Safe Search results page, it contains 'search-results.com'.
Target	Si vous regardez l'URL d'une page de résultats de Recherche sécurisée Norton, elle contient 'search-results.com'.
	Baseline → 2 **+UGC** → 3 **Ref** → 3
Source	cgoldman wrote:
Target	le cgoldman s'est enregistré :
	Extended → 5 **+UGC** → 2 **Ref** → 2

Table 5: Example of segments where the correct comprehensibility class is predicted using UGC features.

can be biased towards the majority class, the QE task appears to be more difficult with a balanced dataset with a high standard deviation.

The best-performing feature type varies amongst the MT systems. For instance, the features built on the domain-specific translation memories (LMs and n-gram counts) bring more useful information when estimating the translation comprehensibility of the MS translations. It is possible that this is happening because the MOSES and SYSTRAN systems were trained on domain-specific data while the MS system was not. Domain-specific features may be particularly helpful in estimating the quality of the output of a general-purpose, non-domain-tuned MT system.

Although the MS translations represent the most difficult set for the QE task, they are the best translations according to BLEU score (0.39 compared to 0.37 for MOSES and 0.35 for SYSTRAN). However it is not possible to conclude from this that there is a negative correlation between MT and QE performance since there are only three MT systems and the differences between them are small. Whether or not this points to a general trend requires further experimentation with other QE datasets, feature sets and MT systems.

	base.	win	full	+ full sel.	+ type sel.
MAE	0.69	0.61	0.67	0.67	0.67
RMSE	0.82	0.75	0.84	0.84	0.83

Table 6: Comparison between the baseline, the shared task winner and our approach on the WMT12 QE dataset.

To test the portability of our feature set and feature selection methods, we evaluate them on the WMT 2012 shared task dataset. The feature set contains the same features as the *Extended* set, apart from the ones which could not be extracted from the training data provided by the shared task organisers, such as forum and Symantec TM LMs and n-gram features. We report the results in Table 6. We do not outperform the baseline features in terms of RMSE but we do with MAE. Feature selection does not bring any improvement in MAE, but RMSE is slightly improved when selection is carried out at the level of individual feature type. Our system lags behind the top-ranked system (Soricut et al., 2012) – more feature selection experimentation is required in order to narrow this gap. It would also be interesting to see how the top-ranked system performs on our UGC dataset.

7 Conclusion

We have conducted a series of quality estimation experiments on English-French user-generated content, estimating both translation comprehensibility and fidelity, and training systems on the output of three individual MT systems. The experiments show that the information brought by a type of feature can be more or less useful depending on the MT system used. We show that the baseline suggested by the WMT12 QE shared task organisers leads to respectable results on user-generated content but we also show that there is sometimes some modest benefit to be found in extending this feature set. The features that are designed specifically to take into account that our data is user-generated content did not perform as well as other new features. However, we cannot conclude from this that modelling the forum characteristics of our data is unnecessary since the LM features trained on forum text perform well. In the future we plan to apply QE at the level of forum post rather than segment.

Acknowledgments

The research reported in this paper has been supported by the Irish Research Council Enterprise Partnership Scheme (EPSPG/2011/102 and EPSPD/2011/135) and the computing infrastructure of the Centre for Next Generation Localisation (`www.cngl.ie`) at Dublin City University.

References

Joshua S. Albrecht and Rebecca Hwa. 2007. Regression for Sentence-Level MT Evaluation with Pseudo References. In *ACL*, pages 296–303.

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier, Andy Way, and Josef van Genabith. 2012. Domain Adaptation in SMT of User-Generated Forum Content Guided by OOV Word Reduction: Normalization and/or Supplementary Data? In *EAMT*, pages 169–176.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *COLING/ACL Workshop on Interactive presentation sessions*, pages 69–72.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence Estimation for Machine Translation. In *JHU/CLSP Summer Workshop Final Report*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *WMT*, pages 1–44.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *WMT*, pages 10–51.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Jacques Duchateau, Kris Demuynck, and Patrick Wambacq. 2002. Confidence Scoring Based on Backward Language Models. In *ICASSP*, pages 221–224.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-Level MT Evaluation Without Reference Translations: Beyond Language Modeling. In *EAMT*, pages 103–111.

Simona Gandrabur and George Foster. 2003. Confidence Estimation for Translation Prediction. In *CoNLL*, pages 95–102.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *ACL*, pages 423–430.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL*, pages 177–180.

Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *ACL*, pages 440–447.

Christopher Quirk. 2004. Training a Sentence-Level Machine Translation Confidence Measure. In *LREC*, pages 825–828.

Sylvain Raybaud, Caroline Lavecchia, David Langlois, and Kamel Smaıli. 2009. Word-and Sentence-Level Confidence Measures for Machine Translation. In *EAMT*, pages 104–111.

Sylvain Raybaud, David Langlois, and Kamel Smaïli. 2011. "This Sentence is Wrong." Detecting Errors in Machine-Translated Sentences. *Machine Translation*, pages 1–34.

Johann Roturier and Anthony Bensadoun. 2011. Evaluation of MT Systems to Translate User Generated Content. In *MT Summit*, pages 244–251.

Raphael Rubino, Jennifer Foster, Joachim Wagner, Johann Roturier, Rasoul Samad Zadeh Kaljahi, and Fred Hollowood. 2012. DCU-Symantec Submission for the WMT 2012 Quality Estimation Task. In *WMT*, pages 138–144.

Raphael Rubino, José Guilherme Camargo de Souza, Jennifer Foster, and Lucia Specia. 2013. Topic Models for Translation Quality Estimation for Gisting Purposes. In *MT Summit*.

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

Lucia Specia and Jesús Gimenez. 2010. Combining Confidence Estimation and Reference-Based Metrics for Segment Level MT Evaluation. In *AMTA*.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the Sentencel-Level Quality of Machine Translation Systems. In *EAMT*, pages 28–35.

Nicola Ueffing and Hermann Ney. 2004. Bayes Decision Rules and Confidence Measures for Statistical Machine Translation. *Advances in Natural Language Processing*, pages 70–81.

Nicola Ueffing, Klaus Macherey, and Hermann Ney. 2003. Confidence Measures for Statistical Machine Translation. In *MT Summit*.