Topic Models for Translation Quality Estimation for Gisting Purposes

Raphael Rubino[†], **José G. C. de Souza**[‡], **Jennifer Foster**[†], **Lucia Specia**[⋆]
[†] NCLT, School of Computing, Dublin City University, Ireland
{rrubino,jfoster}@computing.dcu.ie
[‡] Fondazione Bruno Kessler, Italy
desouza@fbk.eu
[⋆] Department of Computer Science, University of Sheffield, United Kingdom
l.specia@sheffield.ac.uk

Abstract

This paper addresses the problem of predicting how adequate a machine translation is for gisting purposes. It focuses on the contribution of lexicalised features based on different types of topic models, as we believe these features are more robust than those used in previous work, which depend on linguistic processors that are often unreliable on automatic translations. Experiments with a number of datasets show promising results: the use of topic models outperforms the state-of-the-art approaches by a large margin in all datasets annotated for adequacy.

1 Introduction

Quality estimation (QE) for machine translation (MT) is an area concerned with predicting a quality indicator for an automatically translated text without referring to human translations (the so-called reference translations typical of most MT evaluation metrics) (Blatz et al., 2004; Specia et al., 2009).

The widespread use of MT in the translation industry has strongly motivated work in this area. As a consequence, the majority of existing work focuses on predicting some form of post-editing effort to help professional translators (Section 2). However, one equally appealing application is that of estimating the *adequacy* of translations for gisting purposes. An indicator of such a type is particularly relevant in contexts where the reader does not know the source language.

QE is generally addressed as a machine learning task. Intuitively, it is expected that features used to capture general aspects of quality are different from features that define adequacy. Previous work on adequacy estimation has focused on linguistic features contrasting the source and translation texts (Specia et al., 2011; Mehdad et al., 2012), e.g. the proportion of overlapping typed dependency relations in the source and target sentences with arguments that align to each other (based on word-alignment information). While these can provide interesting indicators, they are often very sparse and noisy. Sparsity happens because many of these features do not apply to most sentences, such as features comparing named entities in the source and target sentences. A significant amount of noise can come from the fact that linguistic processors, such as syntactic parsers and named entity recognisers, need to be applied to potentially low-quality translations, and therefore their outcome becomes less reliable. In addition, these indicators rely on external resources that are not available to many languages.

We propose to use topic modelling (TM) features for this problem. TM features significantly differ from those used in previous work in that they focus on important (content) words in the source and target texts, as opposed to more abstract linguistic relationships between these words. We believe these are more robust as they do not depend on further analysis, and that they can be made less sparse through the exploitation of models with different dimensionalities and the use of distance metrics between topic distributions as opposed to topic distributions themselves. A challenge we face is how to model topics in a bilingual setting. We exploit two variants of TMs for that: Polylingual Topic Model (Mimno et al., 2009) and

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 295–302.

a joint Latent Dirichlet Allocation approach (Blei et al., 2003), and a few variants of features based on these models, including the word distribution themselves and distance metrics between source and target distributions (Section 3).

We experiment with three families of datasets: two annotated for adequacy, containing newswire and user-generated content (from a product forum), and one news dataset annotated for post-editing effort (Section 4). We show that TM features are more effective for both adequacy-annotated types of datasets (Section 5).

2 Related Work

Most research work on QE for machine translation is focused on feature engineering and feature selection, with some recent work on devising more reliable and less subjective quality labels (Specia and Farzindar, 2010; Specia, 2011). Blatz et al. (2004) present the first comprehensive study on QE for MT: 91 features were proposed and used to train predictors based on an automatic measure (e.g. NIST (Doddington, 2002)) as the quality label.

Examples of successful cases of QE include improving post-editing efficiency by filtering out low quality segments which would require more effort or time to correct than translating from scratch (Specia, 2011), selecting high quality segments to be published as they are, without post-editing (Soricut and Echihabi, 2010), selecting a translation from either an MT system or a translation memory for post-editing (He et al., 2010), selecting the best translation from multiple MT systems (Specia et al., 2010), and highlighting subsegments that need revision (Bach et al., 2011). For an overview on various feature sets and machine learning algorithms, we refer the reader to the recent shared-task on the topic (Callison-Burch et al., 2012).

Most QE work focuses on estimating a score that indicates overall quality having professional translators as intended user, e.g. post-editing effort. Little work has been done for other applications of MT, such as gisting. One notable exception is the work by Specia et al. (2011), where adequacy scores for Arabic-English translations are predicted. The feature set used include standard features that try to capture general aspects of quality (such as language models of translations), and a range of linguistically motivated features based on part-of-speech tags, dependency trees, and named entities, extracted for both source and target sentences individually, and also contrasting source and target, e.g. ratio of named entities of a given type in the source and target sentences.

A second example is the work by Mehdad et al. (2012), which is tested on datasets annotated for adequacy by volunteers as part of the WMT evaluation campaign, and on datasets annotated with more general quality labels for post-editing effort. The approach uses the framework of cross-lingual textual entailment recognition to address adequacy evaluation. Bi-directional entailment between source and target is considered as evidence of translation adequacy. The framework uses a combination of surface, syntactic and semantic features similar to those used in (Specia et al., 2011), extracted from both source and target sentences, e.g. common dependency relations in source and target sentences.

Both previous approaches for adequacy estimation severely suffer from data sparsity while attempting to model contrastive linguistic information between source and target sentences. As a consequence, the reported results are poor, sometimes even below simple baselines such as the majority class on the training data. None of the previous work uses lexicalised features or topic models built based on those features for adequacy estimation. As we will discuss in the next section, Rubino et al. (2012) used topic models as part of a larger feature set to estimate post-editing effort. However, the contribution of this information source was not tested.

3 Topic Models for Quality Estimation

The first study on using topic modelling for QE was conducted in (Rubino et al., 2012) as part of the WMT12 QE shared task to estimate the post-editing effort of news texts translated from English to Spanish. A joint LDA approach was used. We expand on that work by exploring two types of bilingual topic models and defining a number of variants of features based on source and target (translation) distributions over the dimensions of a topic space. In addition to a joint LDA approach based on the classic LDA model introduced in (Blei et al., 2003), we build a Polylingual Topic Model as presented in (Mimno et al., 2009).

These two models are very different. For the joint LDA approach, the two sides of a large parallel corpus are concatenated at the sentence level, resulting in one corpus containing each source sentence and its translation in the same line. Therefore, source and target languages become indistinguishable. From this, one topic model is built, where each dimension contains the vocabulary in the two languages. Conversely, for the Polylingual model two dimension-aligned monolingual topic models are built, each containing the vocabulary of the corresponding language. Both topic models are built using the Mallet toolkit (McCallum, 2002).

In order to evaluate the use of topic modelling for machine translation QE, we use the classic approach based on feature extraction from source and target sentences, followed by a machine learning step. For the feature extraction step, we consider four different configurations for each topic modelling approach, based on the inferred topic distributions of the source sentences and their translations noted $p(w_n|z_n, \beta)$, for the words w_n of a sentence s, conditioned on the topic z_n with β as the Dirichlet prior on the per-topic sentence distribution. We consider two types of topic-based features to extract:

- the distributions of source sentences and their translations over the topics;
- the source and target distributions divergence and distance measured using five metrics (see below);

To compute the distance between source and target distributions over topics, we consider the classic metrics used in Euclidean geometry, assuming that the topic distributions are represented in an inner product space. For instance, the Euclidean distance is an easy way to measure the length of the segment connecting two points. In an n-dimensional space, the Euclidean distance between two vectors u and v is given by (1):

$$euclidean(u, v) = \sqrt{\sum_{i=1}^{n} |u_i - v_i|^2} \quad (1)$$

Another widely used metric in language processing for measuring the distance between two n-dimensional vectors is the cosine distance, incorporating the inner product in the similarity computation. In our case, two vectors are compared, where each dimension is the probability of a word for a given topic. Thus, we assume that the cosine distance indicates the source and target topic distributions orientations. In an n-dimensional topic space, the cosine distance between two vectors u and v is given by (2):

$$cos(u, v) = \frac{\sum_{i=1}^{n} u_i \times v_i}{\sqrt{\sum_{i=1}^{n} (u_i)^2} \times \sqrt{\sum_{i=1}^{n} (v_i)^2}} \quad (2)$$

Instead of measuring the distance between two n-dimensional sets of points, it is also possible to compute the sum of the absolute differences of their coordinates. These metrics, usually categorised as the Minkowski family of distance metrics, are inspired by a grid, *city-like*, organisation of the space. Usually referred as rectilinear or Manhattan distance, the city-block distance is directly inspired by the Euclidean distance (Krause, 1975) and has shown interesting results when applied to language processing tasks, such as context-based terminology translation (Laroche and Langlais, 2010). The city-block distance between two n-dimensional vectors u and v is given by (3):

$$cityblock(u, v) = \sum_{i=1}^{n} |u_i - v_i| \quad (3)$$

These three metrics allow us to compute the distance between source and target distributions assuming that they are represented in an Euclidean space. To avoid this constraint, we use two other measures in this study, based on probabilistic uncertainty as introduced by Shannon's work (Shannon, 1948). With the measure of relative entropy, an asymmetric way of comparing two distributions suggested in (Kullback and Leibler, 1951) is given by (4):

$$KL(u, v) = \sum_{i=1}^{n} u_i \ln \frac{u_i}{v_i} \quad (4)$$

This measure, also referred to as information deviation, is the basis of many variants, like the *J* or *K* divergences. These measures, all asymmetric, have their symmetric variants, like the *Topsøe* or the *Jensen-Shannon* divergences. The latter one is a symmetric version of the *K* divergence, given by (5):

$$JS(u, v) = \frac{1}{2} \sum_{i=1}^{n} u_i \ln \left(\frac{2u_i}{u_i + v_i} \right) + \sum_{i=1}^{n} v_i \ln \left(\frac{2v_i}{u_i + v_i} \right) \quad (5)$$

The five measures presented in this section are computed and used as features for each source-translation pair. We assume that these 5-dimensions feature vectors can help to capture translation adequacy with lower dimensions compared to the source and translation distributions over the topics. This latter set contains as much features as the number of dimensions in the topic space and thus lead to sparse or noisy features. We decided to use only the topic distance and divergence measures as features in all experiments presented in this paper.

4 Experimental Settings

Three groups of QE datasets are considered in our experiments, two annotated in terms of translation adequacy (for gisting purposes) and the third focusing on post-editing effort. As the language pairs involved are different for each dataset, we use different parallel data to build the topic models.

4.1 Datasets

Arabic-English Data The Arabic-English dataset consists of newswire data from the DARPA GALE project (MT08-nw, GALE09-dev-nw, GALE10-dev-nw) (Specia et al., 2011). More specifically, the dataset contains $2,585$ Arabic sentences and their translation produced by two state-of-the-art MOSES-like phrase-based SMT systems (SMT-1 and SMT-2) and annotated by a professional translator with adequacy scores. The adequacy scores range from 1 (completely inadequate) to 4 (highly adequate), with intermediate categories indicating poorly (2) or fairly (3) adequate translations. We use 80% of these sentences to train the QE models, and the remaining to test them, over three different random splits of the data. To build the topic models we use a concatenation of all Arabic-English newswire parallel data provided by LDC and the Arabic-English UN data,[1] totalling $\sim$ 6.4M translation pairs after removing sentences longer than 80 words. This was virtually the same parallel corpus used to build the SMT systems.

User-Generated Data The user-generated content is composed of 694 sentences taken from an English IT-related online forum, translated into French by three automatic translators considered as *black-box* systems: MOSES, BING and SYSTRAN. The MOSES system is a standard phrase-based SMT system trained using the Moses (Koehn et al., 2007) and IRSTLM (Federico et al., 2008) toolkits and optimised on a development set against BLEU (Papineni et al., 2002) using MERT (Och, 2003). A trigram Language Model (LM) was built using Witten-Bell smoothing. This system was trained using in-domain translation memories (up to $\sim$ 1.6M translation units) plus $\sim$ 1M translation units from the Computer Software domain (obtained from the TAUS Data Association[2]) for the translation model, as well as additional monolingual forum data (up to 20K French sentences) for the LM. The BING system is a freely available generic SMT system, Bing Translator,[3] accessed through the second version of their API. Finally, the last system is the Systran Enterprise Server version 6, customised with the use of a domain specific 10K+ dictionary entries.

Each translation is evaluated by a professional translator using two possible labels: 0 if the translation does not preserve the meaning of the source sentence and 1 if the meaning is preserved. The final dataset contains, for each of the three translations generated by the three MT systems, 694 source segments, 694 translated segments, and one adequacy score. From this dataset, 500 segments are used to train the QE models and 194 segments are held out for evaluation purposes. More information about this dataset can be found in (Roturier and Bensadoun, 2011). To build the topic models for the adequacy features, we use the in-domain Translation Memories used to train the MOSES system.

WMT12 QE Data The WMT12 QE dataset (Callison-Burch et al., 2012) is composed of $2,254$ English sentences translated into Spanish by a MOSES phrase-based system and evaluated by three professional translators in terms of post-editing effort on a 1 (highest) to 5 scale (lowest). The three scores per segment pair are weighted and averaged in order to obtain one continuous score ($\in [1;5]$). From this dataset, $1,832$ segments are used to train the QE models while 422 segments are used as a test set. To build the topic models, we use the parallel corpora used by the MOSES system which generated the Spanish translations. This

[1] http://www.uncorpora.org/

[2] http://www.tausdata.org/

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

corpus contains the concatenation of Europarl v5 and the News Commentary corpus from WMT10 translation task in English and Spanish ($\sim$ 1.7M translation pairs) (Callison-Burch et al., 2010).

4.2 Additional Features

In addition to the TM features described in Section 3, for all datasets, for comparison we consider a **baseline** set of 17 features that performed well across languages in previous work and were used as the official baseline in the WMT12 QE task (Callison-Burch et al., 2012):
– number of tokens in the source & target sentences;
– average source token length;
– average number of occurrences of the target word within the target sentence;
– number of punctuation marks in source and target sentences;
– language model (LM) probability of source and target sentences using 3-gram LMs built from the source/target sides of SMT training corpus;
– average number of translations per source word as given by IBM 1 model thresholded such that $P(t|s) > 0.2$;
– same as above with $P(t|s) > 0.01$ weighted by the inverse frequency of each word in the source side of the SMT training corpus; – percentage of unigrams, bigrams and trigrams in frequency quartiles 1 (lower frequency words) and 4 (higher frequency words) in the source side of the SMT training corpus;
– percentage of unigrams in the source sentence seen in the source side of the SMT training corpus.

In addition, for the Ar-En dataset, we have access to a larger sets of features (including the 17 baseline features): **black-box** (BB) and **glass-box** (GB) features, where the latter depend on internal information from the MT systems that produced the translations. In total 122 BB features containing language-specific variants are used for both Ar-En datasets, with additional 39 (SMT-1) or 48 (SMT-2) GB features (different SMT systems provide different features). For a description of these features, see (Specia et al., 2011). In our experiments we also test combinations of these baseline, BB and GB features with the proposed topic model features. Baseline, BB and GB features were extracted using the open source toolkit QuEst.[4]

[4] http://www.quest.dcs.shef.ac.uk

4.3 Learning Algorithms

In all the experiments presented in this paper, we used the LIBSVM (Chang and Lin, 2011) implementation of Support Vector Machine (SVM) to build the regression and classification models. For the Arabic-English and WMT12 datasets, the regression models were trained using the ϵ-SVR algorithm. While learning regression models is the most common strategy for QE, learning a classifier is more appropriate for the binary labels in the user-generated data. For this dataset, a classification model was trained with the c-SVC algorithm. For these two SVM algorithms, a radial basis function (RBF) kernel is used and its parameters are optimised by grid-search on the training data, performing a 5-fold cross-validation for each set of parameters, keeping the best parameters according to Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for the regression models and accuracy for the classification model.

4.4 Evaluation Metrics

We evaluate the results in terms of two error metrics for the regression tasks (for the En-Ar and WMT12 QE datasets), **MAE** and **RMSE**:

$$MAE = \frac{\sum_{i=1}^{N} |H(s_i) - V(s_i)|}{N}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (H(s_i) - V(s_i))^2}{N}}$$

In both MAE and RMSE, H is the prediction computed by the system and V is the true value obtained from labelled data. N is the total number of instances in the dataset. For the regression tasks, we also present a common baseline: predicting the *mean* of the scores in the training data. In other words, we assign the mean of the training data labels to all test instances and measure error.

For the binary classification task (user-generated dataset), we evaluate our approach by computing the precision score on class 1, *i.e.*, when the translation keeps the meaning of the source sentence, thus focusing on the usability of topic model features for adequacy estimation.

5 Experiments

5.1 Arabic-English Data

Several topic model configurations were considered, from 50 to 400 topics for both the joint

LDA and the Polylingual approaches, and the best results were obtained with topic model features (noted TM) extracted from a 200-topic Polylingual model. These results are shown in Table 1, where MAE and RMSE scores are computed by averaging the results obtained for the three folds.

Feature set	SMT	MAE	RMSE
Mean	1	0.6050	0.7733
	2	0.5445	0.7586
Baseline	1	0.5680	0.7187
	2	0.5497	0.7309
Baseline+TM	1	**0.5507**	0.7234
	2	0.5629	0.7002
BB+GB	1	0.5649	0.7082
	2	0.5496	0.7051
BB+GB+TM	1	0.5619	0.7007
	2	0.5397	0.6716
BB+GB+TM "and"	1	0.5589	**0.6963**
	2	0.5448	0.6875
BB+GB+TM "maj"	1	0.5655	0.7048
	2	**0.5380**	**0.6705**

Table 1: MAE and RMSE results when estimating adequacy for the Arabic-English dataset using a Polylingual topic model with 200 topics.

Even though the mean baseline proved to be very hard to beat for the SMT-2 system, we notice that the combination BB+GB+TM improves the RMSE when applying the models for both systems. Given the large number of features in this dataset, we also performed feature selection. The best RMSE score for SMT-1 is obtained with feature selection using Randomized Lasso with an "and" fold combination (intersection of the features groups selected in each fold). Similarly, the best RMSE score for SMT-2 is obtained using feature selection but with a majority vote among the three folds. We observed that at least one topic-based feature is always kept in the selected feature sets. These results show that adding the topic model features to the baseline set outperforms all the other configurations in terms of MAE for SMT-1. A reduction of 0.0173 is observed with this feature set compared to the baseline.

5.2 User-Generated Data

Predicting the translation adequacy of the user-generated dataset was done for each translation system individually. A binary classification setup was designed and we evaluate our approach by measuring the precision on class 1 (meaning preserving). For both topic modelling approaches, three configurations are considered in terms of topic space dimensionality: 10, 50 and 100 topics. The extracted distances and divergences between source and target distributions over topics are directly used as features and also combined with the 17 baseline features to train the classification models. Results are presented in Table 2.

Feature set		MOSES	SYSTRAN	BING
Baseline		0.711	0.569	0.709
Joint	10top.	0.583	**0.667**	0.528
	50top.	0.607	0.652	0.567
	100top.	**0.850**	0.625	0.636
	Base+10top.	0.750	0.522	0.686
	Base+50top.	0.625	0.547	**0.791**
	Base+100top.	0.800	0.654	0.740
Poly.	10top.	0.711	0.633	0.719
	50top.	0.613	0.593	0.621
	100top.	0.600	0.546	0.667
	Base+10top.	0.781	0.571	0.706
	Base+50top.	0.657	0.586	0.695
	Base+100top.	0.641	0.623	0.691

Table 2: Precision results for the user-generated data using 10 to 100 dimensions topic models.

Unlike the results obtained on the Arabic-English data, the joint LDA approach leads to the best results for the user-generated data. The precision scores for the three translation systems are improved by using topic model features compared to the baseline. For MOSES and SYSTRAN, the best results are obtained with the topic model features without any additional features, for 100 and 10 topics respectively. It appears that increasing the number of dimensions of the topic space improves the classification precision for MOSES when standalone topic features are used, while the opposite is observed with SYSTRAN. For BING, the additional 17 baseline features help improve over the topic features alone. Some examples of correctly predicted adequacy classes are presented in Table 3.

5.3 WMT12 QE Data

For the WMT12 QE data, as with the user-generated data, both Joint LDA and Polylingual topic models features are evaluated. An important aspect of this dataset is related to the scores we attempt to estimate: they do not focus on translation adequacy, but rather on overall quality, taking into consideration the post-editing effort required to reach an acceptable translation. The results are presented in Table 4, where the baseline scores are the ones reported in the WMT12 QE shared task.

Source	it might be google :
Target	ce pourrait être google :
Baseline	0
Joint	1
Source	thanks for more information than i had any idea was available...
Target	merci pour plus d' informations que j' avais une idée était disponible...
Baseline	1
Polylingual	0
Source	much thanks and appreciation
Target	appréciation et merci encore
Baseline	1
Polylingual	0
Source	proprietary information subject to a confidentiality agreement .
Target	l' information exlusive sujet à une convention de confidentialité .
Baseline	1
Joint	0
Source	this file-by-file thing is going to take 3 months .
Target	ce fichier chose va prendre 3 mois .
Baseline	1
Polylingual	0
Source	the icon raises some security issues .
Target	l' icône soulève quelques problèmes de sécurité .
Baseline	0
Joint	1

Table 3: Source and target UGC segments when topic-based features lead to a correct binary classification, compared to the baseline.

For both the joint and Polylingual topic models, it is possible to notice a consistent improvement in performance when increasing the number of topics. The best performance with joint LDA is achieved with 50 topics when combined with the baseline features. With the Polylingual topic model, the best MAE is achieved with 100 topics combined with the baseline whereas 10 topics plus the baseline gives the best RMSE. However, the RMSE scores obtained with the 17 baseline features outperforms all the topic model configurations evaluated in this set of experiments. This phenomenon indicates that the prediction errors are larger when using topic model features compared to the baseline. They may also suggest that topic model features are more appropriate for adequacy estimation, as we hypothesised.

6 Conclusion

This paper presents a novel approach to estimate the translation adequacy based on topic model features. Two kinds of topic models are evaluated, a Joint LDA Model and a Polylingual Model.

Feature set		MAE	RMSE
Mean		0.8279	0.9899
Baseline		0.69	**0.82**
Joint	10top.	0.8066	0.9721
	50top.	0.8147	0.9805
	100top.	0.7851	0.9601
	Base+10 top.	0.6892	0.8519
	Base+50 top.	**0.6783**	0.8404
	Base+100 top.	0.6930	0.8574
Poly.	10top.	0.8113	0.9739
	50top.	0.7845	0.9481
	100top.	0.7773	0.9456
	Base+10 top.	0.7029	0.8650
	Base+50 top.	0.7095	0.8866
	Base+100 top.	0.6943	0.8709

Table 4: MAE and RMSE results when estimating the post-editing effort for the WMT12 QE dataset using 10 to 100 dimensions topic models.

The evaluation was conducted on three types of datasets: Arabic-English newswire data annotated in terms of adequacy, English-French data taken from an online IT forum where the content is user-generated, and English-Spanish news data with a focus on post-editing effort estimation.

We investigate the impact of topic model features through a systematic evaluation of the two topic modelling approaches with different configurations in terms of topic space dimensionality, feature combinations and feature selection. Overall, different configurations of topic models were found to perform better on different datasets. Nevertheless, some general conclusions can be made: the results obtained indicate that the distance and divergence between source and target sentence distributions over topics are very effective as features to estimate translation adequacy. Experiments on the Arabic-English and the user-generated datasets show that topic model features outperform strong baselines.

As future work, we plan to study the impact of different pre-processing techniques on the training data used to build the topic models, including stemming and stop-word filtering. We also want to conduct a more in depth analysis of the quality estimation results obtained on the user-generated content translated by diverse machine translation systems, as a variable number of topics leads to the best classification results.

Acknowledgements

The research reported in this paper has been partly supported by the Research Ireland Enter-

prise Partnership Scheme (EPSPG/2011/102 and EPSPD/2011/135). We would like to thank Dr. Fred Hollowood and Dr. Johann Roturier for providing us with the user-generated content dataset.

References

Bach, Nguyen, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: a method for measuring machine translation confidence. In *ACL11*, pages 211–219.

Blatz, John, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence Estimation for Machine Translation. In *Coling04*, pages 315–321.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

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

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *WMT12*, pages 10–51.

Chang, Chih-Chung and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

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

Federico, Marcello, Nicola Bertoldi, and Mauro Cettolo. 2008. Irstlm: an open source toolkit for handling large scale language models. In *Interspeech*, pages 1618–1621.

He, Yifan, Yanjun Ma, Josef van Genabith, and Andy Way. 2010. Bridging smt and tm with translation recommendation. In *ACL2010*, pages 622–630.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*, pages 177–180.

Krause, Eugene F. 1975. *Taxicab geometry*. Addison Wesley Publishing Company.

Kullback, Solomon and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Laroche, Audrey and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *COLING*, pages 617–625.

McCallum, Andrew Kachites. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Mehdad, Yashar, Matteo Negri, and Marcello Federico. 2012. Match without a referee: evaluating mt adequacy without reference translations. In *WMT*, pages 171–180.

Mimno, David, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 2009. Polylingual topic models. In *EMNLP*, pages 880–889.

Och, F.J. 2003. Minimum error rate training in statistical machine translation. In *ACL*, volume 1, pages 160–167.

Papineni, K., S. Roukos, T. Ward, and W.J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Roturier, Johann and Anthony Bensadoun. 2011. Evaluation of mt systems to translate user generated content. In *MT Summit XIII*, pages 244–251.

Rubino, Raphael, Jennifer Foster, Joachim Wagner, Johann Roturier, Rasul Samad Zadeh Kaljahi, and Fred Hollowood. 2012. Dcu-symantec submission for the wmt 2012 quality estimation task. In *WMT*, pages 138–144.

Shannon, Claude E. 1948. The bell system technical journal–vol. 27. *July, October*.

Soricut, Radu and Abdessamad Echihabi. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *ACL*, pages 612–621.

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

Specia, Lucia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *EAMT09*, pages 28–37, Barcelona.

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

Specia, Lucia, Najeh Hajlaoui, Catalina Hallett, and Wilker Aziz. 2011. Predicting machine translation adequacy. In *MT Summit XIII*.

Specia, Lucia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *EAMT11*, pages 73–80.