MT Quality Estimation for E-Commerce Data

José G. C. de Souza desouza@fbk.eu
Marcello Federico federico@fbk.eu
Fondazione Bruno Kessler, Trento, Italy

Hassan Sawaf hsawaf@ebay.com
Human Language Technology unit, eBay Inc., San Jose, USA

Abstract

In this paper we present a system that automatically estimates the quality of machine translated segments of e-commerce data without relying on reference translations. Such approach can be used to estimate the quality of machine translated text in scenarios in which references are not available. Quality estimation (QE) can be applied to select translations to be post-edited, choose the best translation from a pool of machine translation (MT) outputs, or help in the process of revision of translations, among other applications. Our approach is based on supervised machine learning algorithms that are used to train models that predict post-editing effort. The post-editing effort is measured according to the translation error rate (TER) between machine translated segments against their human post-edits. The predictions are computed at the segment level and can be easily extended to any kind of text ranging from item titles to item descriptions. In addition, our approach can be applied to different kinds of e-commerce data (e.g. different categories of products). Our models explore linguistic information regarding the complexity of the source sentence, the fluency of the translation in the target language and the adequacy of the translation with respect to its source sentence. In particular, we show that the use of named entity recognition systems as one source of linguistic information substantially improves the models' performance. In order to evaluate the efficiency of our approach, we evaluate the quality scores assigned by the QE system (predicted TER) against the human post-editions (real TER) using the Pearson correlation coefficient.

1 Introduction

Approaches to machine translation (MT) quality estimation (QE) are used in situations in which a quality score about the translation is required but no references translations are available. In MT QE, automatically translated sentences have their quality estimated without using references. Such scenarios include supporting the work of translators in a CAT scenario (Turchi et al., 2015), informing readers of the translation whether the translation is reliable or not (Turchi et al., 2012), selection of the best translation generated by a pool of MT systems (Specia et al., 2010), or filtering out low-quality translation suggestions that should be rewritten from scratch (Specia et al., 2009).

QE is usually cast as a classification, regression or ranking problem that is modelled using supervised learning techniques. The different forms of supervision used to train the models imply different ways of perceiving the quality of a translation. The choice of the supervision label depends on the envisaged application scenario. For example, for regression and ranking, previous work employed either the time required to post-edit the translations or the minimum number of modifications required to make the translation acceptable as measured by the human

translation error rate (HTER[1], see Snover et al. (2009)). Another required information that must be defined a priori is the kind of linguistic cues that are going to be used to predict quality. Such indicators are extracted from the source and the translated sentence and aim to serve as a proxy for the complexity of translating the source sentence, the fluency of the translated sentence and the adequacy of the translation in function of the source.

In this work we present the first approach to MT QE geared towards e-commerce user-generated data. Our challenge is two-fold: (i) the data have been generated by many users and therefore are not necessarily composed of grammatically well-formed sentences, and (ii) they belong to a domain composed of very diverse topics (read different categories of products). We propose new features designed to deal with the characteristics of these data and evaluate our models against post-edits produced by humans.

2 Background

eBay is a marketplace platform in which sellers can advertise items and buyers can search for items, electronically bid and eventually buy them. To enable trade between buyers and sellers with different languages, at least four types of texts need to be translated: queries, item titles, descriptions, and item specifics. Machine translation has been recently introduced in eBay's platform with the objective of fostering cross-border trade between sellers and buyers that speak different languages (Guha and Heger, 2014). In this work we predict the quality of translation of item titles, which are concise and usually very informative descriptions of items put on sale. One item title example is given below:

> *Universal 12000mAh Backup External Battery USB Power Bank Charger for Cell Phone*

It specifies several characteristics of a product ranging from more generic information (i.e. "Backup External Battery") to more specific characteristics (i.e. 12000 mAh). Common challenges in the translation of eBay's user generated content in general, and of titles (Sanchez and Badeka, 2014) are the correct rendering of proper names and the translation of words which can have multiple senses, depending on the context in which they appear. Furthermore, words can appear in a relative free-order in the title without damaging its meaning. This presents a challenge for MT QE models because they assume that source sentences are well-formed and grammatical in the source language.

3 Related Work

Most of the work for MT QE has been developed using well-formed and grammatical sentences belonging to different domains such as legal (transcription of political speeches), or news-wire texts covering different topics (Callison-Burch et al., 2012; Bojar et al., 2013, 2014). Likewise, all the features designed in previous work assume that source sentences are grammatical and that the MT systems were trained over parallel data with fluent and well-formed segments.

To the best of our knowledge, the first MT QE approach to consider user-generated data was presented by Rubino et al. (2013a). In this work, the authors present regression and classification models trained and evaluated on two different language pairs and two different domains. In particular, they developed a QE classification model for English-French information technology forums data described in Roturier and Bensadoun (2011). The approach explores features based on topic models that focus on the adequacy aspect of the translations (i.e. check whether the meaning of the source sentence is present in its translation).

[1]The translation error rate is calculated as the number of edits (word insertions, deletions, substitutions, and shifts) divided by the number of words in the reference. Lower HTER values indicate better translations.

The same English-French dataset is used by Rubino et al. (2013b) to develop QE classification models with features that were tailored to be more discriminative on user-generated data. The features specific to user-generated data explore inconsistent use of character case, non-standard punctuation, spelling mistakes and sentence splitting problems. [talk about performance and best feature sets]

Previous work differs from our work in two main aspects: quality label and data domain. The quality labels used by Rubino et al. (2013a,b) describe whether a translation is adequate or not. Binary classification models are developed aiming to predict the adequacy of translations. In this work, instead, we focus on predicting post-edition effort as a proxy for quality by training regression models. Furthermore, the data domain of previous work is information technology forums whereas our focus is on e-commerce data that spans several products in different categories.

4 MT QE for E-Commerce

In this section we describe our approach to MT QE for e-commerce data. We first describe the features we extract from both source and translation sentences and then we move to the description of the learning algorithms used for training QE models.

4.1 Features

We use a combination of features that mixes general-purpose linguistic cues with features designed specifically for the kind of data we are dealing with. We assume that the QE system does not have access to the MT system and therefore we do not use any kind of feature extracted from the MT system translation process. Such assumption allows the features presented here to be used with any MT system.

4.1.1 Domain-independent features

We extract a set of 79 domain-independent features implemented in the QuEst feature extractor framework Specia et al. (2013). These features have been proposed in previous work for MT QE and span three translation aspects: source complexity, translation fluency, and translation adequacy.

The source complexity refers to the difficulty of translating the source sentence. Longer sentences or sentences with more than one clause tend to be more complex to understand and more difficult to translate. Examples of complexity-oriented features are (computed only in the source sentence):

- number of punctuation marks;
- average token length;
- number of tokens.

The translation fluency dimension regards the correct use of grammar in the translation in the target language. The more fluent is the translation generated in the target language, the better the translation is. Examples of fluency features are (computed only in the translation sentence):

- language model log probability for the whole translation sentence;
- language model perplexity for the whole sentence;
- percentage of nouns.

Adequacy-oriented features approximate how much of the meaning of the source sentence is found in its translation. Adequacy features are computed with the source and translation sentences at the same time. Examples of features are:

- ratio of nouns in the source and translation sentences;
- absolute difference between the number of punctuation marks between the source and the translation normalised by translation length.

For a list with descriptions for all 79 features please refer to `http://www.quest.dcs.shef.ac.uk/quest_files/features_blackbox`. These features are referred to as "BB79".

4.1.2 Item title embeddings

For item titles it is more important to have translations that convey the meaning of the source title than fluent discourse in the target language. For this reason, focusing on adequacy features is important: because they can capture the meaning of the title instead of language correctness. Following the recent popularity of word embeddings in the NLP literature, we experiment with `paragraph2vec` (Le and Mikolov, 2014) to obtain embeddings that encode the meaning of a title. The embeddings are trained for both the source and the target side of the available item titles parallel corpus. Both source and target embeddings of a given title are then concatenated and used as features in our regression models. The number of features varies according to the number of dimensions of the embeddings. We experiment with several dimensions and the best results are reported in Section 6. We train the embeddings with the `paragraph2vec` implementation of `gensim`[2] (Řehůřek and Sojka, 2010). These features are referred to as "DM" (distributional memory).

4.1.3 NER-based features

Item titles segments present many word or expressions (formed by more than one word) that are proper names and that should not be translated (such as brand names or technical expressions like USB). A MT QE system could benefit of named entity recognition (NER) system that outputs whether a given token is a named entity. If the token is marked as a named entity it should not be translated and therefore it is possible to check whether the brand or name is preserved in the translation.

We developed a set of three features that verify whether the tokens marked as named entities or "do-not-translate" are in fact not translated in the MT output. The three features are:

- number of "do-not-translate" tokens found in the source sentence;
- number of tokens in the translation segment that exactly match the items marked as "do-not-translation" in the source sentence;
- a ratio between the second and first features above.

Such features rely on a in-house NER system that produces binary tags for each token in a sentence. These features could be considered adequacy-oriented features and are tailored specifically for user-generated e-commerce data that inherent to eBay's platform. These are called "NER" hereafter.

[2] `https://radimrehurek.com/gensim/index.html`

4.2 Learning algorithms

We train our models with two different non-linear ensemble learning algorithms: extremely randomized trees (Geurts et al., 2006) and AdaBoost regression trees (). Both are batch non-linear learning algorithms that also provide the importance of each feature in the final fitted model.

Extremely randomized trees (ET) is a learning algorithm based on an ensemble of decision trees (Breiman et al., 1984). ET is an ensemble of randomized trees in which each decision tree can be parameterized differently. When a tree is built, the node splitting step is done at random by picking the best split among a random subset of the input features. All the trees are grown on the whole training set and the results of the individual trees are combined by averaging their predictions. We explore this model after successful results in MT QE (de Souza et al., 2014a).

The second learning algorithm we use to train our models is AdaBoost Regression (Drucker, 1997) (ADA). This algorithm fits a sequence of weak learners (very small decision trees in our case) on several iterations of modified versions of the data. Training examples receive weights according to the difficulty the model has at predicting them, forcing the algorithm to focus on examples that were incorrectly predicted by previous iterations. The final prediction consists of a weighted majority vote (or sum) of all iterations.

5 Experimental Settings

5.1 MT system

The MT system was trained with in-domain (item titles and item descriptions of e-commerce data) and out of domain parallel data (legal and news-wire texts) for training the word alignments. Translation models were trained using the standard Moses pipeline. Due to the nature of the item titles, no lexicalized reordering model is used. On the target side, trigram language models are trained. The parallel data used to train the system comes from various publicly available collections, proprietary repositories and in-house translated item titles. In particular, in-house translated items, descriptions, and specifics are here considered as in-domain data while all the rest is regarded as out of domain data. A summary of the data used to train the MT system is given in Table 1.

	Train (Out-Domain)	**Train (In-Domain)**
Segments no.	5.28M	336K
Tokens EN	69M	2M
Tokens PT no.	70M	2M

Table 1: Statistics of English-Portuguese parallel data.

5.2 Data

We train and evaluate our models on item titles translated from English to Portuguese with the MT system described in Section 5.1. All the translations were post-edited by professional translators following a conservative post-edition guideline (i.e. the post-editors should focus on the minimum modifications necessary to make the translation acceptable). We worked on a translation job of approximately 11,000 translation units comprising more than 200 product categories. For our experiments, we focus on the three more frequent categories, namely: “Cellphones & Accessories” (CPA), “Cellphones & Smartphones” (CPS) and “Women’s Clothing” (WC).

We compute the HTER scores between the translations and their post-editions[3] for each category. The scores are clipped between 0 and 1. The distributions are very different across the

[3]The HTER scores are computed with the tercom tool available at

	CPA	CPS	WC
Segments no.	854	1,031	834
Tokens EN no.	11,632	11,807	10,118
Tokens PT no.	13,318	13,552	11,686

Table 2: Summary statistics for the data used to train and evaluate the QE models.

different categories, showing a big discrepancy in translation quality. WC has a large mass of segments with HTER close to 1 (which means almost all translations are rewritten from scratch) whereas CPS and CPA are centered around the range that goes from 0.4 to 0.7 HTER. In general, the translation quality for all the categories is low, with most of the segments presenting HTER higher than 0.3. The HTER distributions are shown in Figure 1.

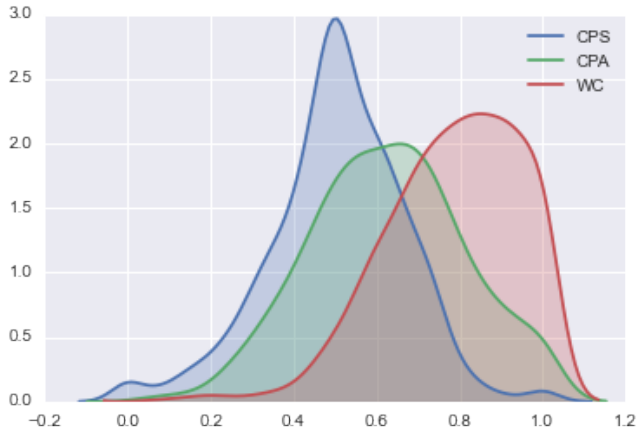

Figure 1: HTER distributions for the segments of the CPA, CPS and WC categories.

5.3 Evaluation metrics

The performance of our regression models is evaluated in terms of two metrics. The first is the mean absolute error (MAE), a standard error measure for regression problems commonly used also for QE Callison-Burch et al. (2012). The MAE is the average of the absolute errors $e_i = |\hat{y}_i - y_i|$, where $\hat{y}_i$ is the prediction of the model and y_i is the true value for the i^{th} instance. As it is an error measure, lower values indicate better performance (↓).

The second is the Pearson correlation, a measure of the linear dependence between two variables. Pearson correlation is defined as the covariance of two variables divided by the product of their standard deviations: $\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$. Higher values indicate better performance (↑).

6 Results and Discussion

In this section we report the results of our models. Hyper-parameters were found with 100 iterations of randomized search (Bergstra and Bengio, 2012) on 5-fold cross-validation over the training data. The final models were trained over the whole training data with the best parameters found during the randomized search procedure.

Results for the three categories evaluated are in Table 3 (CPA), Table 4 (CPS) and Table 5. The first row of each column is a simple baseline that applies the training set HTER mean as

CPA				
	ET		ADA	
Features	**MAE ↓**	**Pearson ↑**	**MAE ↓**	**Pearson ↑**
Mean	15.4	0	15.4	0
BB79	14.29	47.32	13.64	50.27
DM+BB79	14.29	47.6	13.83	46.35
BB79+NER	**13.78**	**50.37**	**13.07**	**55.97**
DM+BB79+NER	13.84	49.87	13.46	51.93

Table 3: Results for the category "Cellphones & Accessories" (CPA).

a prediction for every segment in the test set. This baseline is a lower bound above which our models should perform. Any model with results lower than the "Mean" baseline do not learn anything with the data. All three categories "Mean" baseline presents the highest MAE and correlation equal to zero.

CPS				
	ET		ADA	
Features	**MAE ↓**	**Pearson ↑**	**MAE ↓**	**Pearson ↑**
Mean	12.86	0	12.86	0
BB79	12.42	39.56	11.68	45.57
DM+BB79	12.5	38.72	12.18	41.59
BB79+NER	**12.19**	**44.17**	**11.11**	**53.51**
DM+BB79+NER	12.29	43.42	11.8	49.28

Table 4: Results for the category "Cellphone & Smartphones" (CPS).

Overall, the best feature set is the combination of BB79 and NER for both ET and ADA. For both CPA and CPS this combination presents the best MAE and Pearson correlation. The NER feature set seem to help in particular for the CPA and CPS categories but not so much for WC. The main reason are the characteristics of item titles in the WC category. They contain less brand names and technical concepts about the product and more generic descriptions about clothes, making the named entity information less efficient. Furthermore, the general performance of the QE models for WC is much lower than for the other two categories. The most likely reason is the distribution of HTER labels (Figure 1), which is almost in its entirety composed of bad translations (close to 1 HTER) and very few examples of good translations (close to zero HTER).

WC				
	ET		ADA	
Features	**MAE ↓**	**Pearson ↑**	**MAE ↓**	**Pearson ↑**
Mean	12.99	0	12.99	0
BB79	**12.83**	13.2	13.11	6.75
DM+BB79	12.93	10.04	**12.55**	**11.27**
BB79+NER	12.84	**12.15**	12.93	10.8
DM+BB79+NER	12.93	7.24	12.72	4.14

Table 5: Results for the category "Women's clothing" (WC).

The features based on the title embeddings (DM) do not seem to help the overall performance for predicting post-edition effort. It presents the best results when combined with BB79 and trained with ADA for the WC category, however, the final Pearson correlation is very low if compared with the best models for the other two categories (Pearson correlation of 11.27).

Regarding the learning algorithms, ADA outperforms ET for both CPA and CPS categories. For CPS the results are substantially higher (approximately 1 MAE point and 9 Pearson correlation points). AdaBoost's shortcoming, however, is the time required to train the models. In our experiments, it was as much as 15 times slower than ET.

6.1 Feature analysis

In order to better understand what are the most predictive features for the e-commerce domain we analyze what are the ten most important features according to the models trained with ET for each category. Here we present the features that appear in the intersection of the top-10 most important features for each pair of categories. In the following list, features are sorted by their importance score (for CPA and CPS):

- number of named entities marked as do-not-translate found in the translation;
- number of named entities found in the source sentence (do-not-translate terms);
- ratio of named entities matches found in the translation divided by total number of named entities in the source;
- average number of translations per source word in the sentence (threshold in IBM1: prob > 0.01) weighted by the frequency of each word in the source corpus
- average word frequency: on average, each type (unigram) in the source sentence appears N times in the corpus (in all quartiles);

The most predictive features are the ones related to adequacy and the features developed specifically for eBay's data are the most predictive (NER-based features). For WC, on the contrary, they were not helpful:

- language model log probability of part-of-speech tags in the translation sentence
- language model log probability of the translation sentence

The most predictive features for the WC category are the ones that model fluency in the automatically-generated translation. One interesting avenue of research is to analyse how similar are categories taking into consideration only the features extracted and exploring their similarities and discrepancies in order to build more robust QE models (similarly to de Souza et al. (2014b)).

7 Conclusion

In this paper we presented an approach to MT QE for e-commerce data. We train and evaluate models that predict post-edition effort (HTER) on products from three different categories in the inventory of eBay's marketplace platform. Our models use a combination of domain-independent and domain-specific features and reach approximately 55% correlation when evaluated against post-edit scores produced by professional translators.

As future work, we would like to test our QE system in a localization application scenario. We envisage that such a system could be used to sample segments to be sent for post-edition or to revise post-editions produced by a language service provider (LSP). Many companies and LSPs still rely on a random sampling process that could be improved quality and time-wise by a more informed method that uses MT QE to score translations.

Acknowledgements

The first author received support through a financial gift by eBay Inc. to FBK. The second author received support by the EU H2020 funded MMT project (grant agreement No 645487), by eBay Inc., and by FBK's Mobility programme.

References

Bergstra, J. and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13:281–305.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Eighth Workshop on Statistical Machine Translation*, pages 1–44.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 10–51, Montr{é}al, Canada. Association for Computational Linguistics.

de Souza, J. G. C., González-Rubio, J., Buck, C., Turchi, M., and Negri, M. (2014a). FBK-UPV-UEdin participation in the WMT14 Quality Estimation shared-task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 322–328.

de Souza, J. G. C., Turchi, M., and Negri, M. (2014b). Machine Translation Quality Estimation Across Domains. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.*, pages 409–420.

Drucker, H. (1997). Improving regressors using boosting techniques. In *14th International Conference on Machine Learning*, pages 107–115.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Guha, J. and Heger, C. (2014). Machine Translation for Global E-Commerce on eBay. In *Proceedings of the AMTA*, volume 2: MT Users, pages 31–37.

Le, Q. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *International Conference on Machine Learning - ICML 2014*, 32:1188–1196.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Roturier, J. and Bensadoun, A. (2011). Evaluation of MT Systems to Translate User Generated Content. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 244–251.

Rubino, R., de Souza, J. G. C., and Specia, L. (2013a). Topic Models for Translation Quality Estimation for Gisting Purposes. In *Machine Translation Summit XIV*, pages 295–302.

Rubino, R., Foster, J., Samad Zadeh Kaljahi, R., Roturier, J., and Hollowood, F. (2013b). Estimating the Quality of Translated User-Generated Content. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, number October, pages 1167–1173.

Sanchez, J. and Badeka, T. (2014). Linguistic QA for MT of user-generated content at eBay. In *Proceedings of the AMTA*, volume 2: MT Users, pages 1–24.

Snover, M., Madnani, N., and Dorr, B. (2009). Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, number March, pages 259–268.

Specia, L., Cancedda, N., Dymetman, M., Turchi, M., and Cristianini, N. (2009). Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Annual Conference of the EAMT*, number May, pages 28–35.

Specia, L., Raj, D., and Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

Specia, L., Shah, K., de Souza, J. G. C., and Cohn, T. (2013). QuEst–A translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 79–84.

Turchi, M., Negri, M., and Federico, M. (2015). MT Quality Estimation for Computer-assisted Translation: Does it Really Help ? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 530–535.

Turchi, M., Steinberger, J., and Specia, L. (2012). Relevance Ranking for Translated Texts. In *Proceedings of th 16th International Conference of the European Association for Machine Translation (EAMT)*, number May, pages 153–160.