Bootstrapping Quality Estimation in a live production environment

Joachim Van den Bogaert
CrossLang
Gent, Belgium
joachim@crosslang.com

Bram Vandewalle
CrossLang
Gent, Belgium
bram.vandewalle@crosslang.com

Roko Mijic
CrossLang
Gent, Belgium
roko.mijic@crosslang.com

Abstract

In this paper, we discuss how we bootstrapped Quality Estimation (QE) in a constrained industry setting. No post-edits were at our disposal and only a limited number of annotators was available to provide training data in the form of Post-Edit (PE) effort judgments. We used a minimal approach and applied a simplified annotation procedure. We used as few as 17 baseline features for QE training.

1 Introduction

In this paper, we discuss how we bootstrapped Quality Estimation (QE) – the process of scoring Machine Translation (MT) output without access to a reference translation – for 9 language pairs and 3 domains in a constrained industry setting. No post-edits were at our disposal and only a limited number of annotators was available to provide training data in the form of Post-Edit (PE) effort judgments. We used a minimal approach (Callison-Burch et al., 2009), by annotating only 800 segments per language pair and content type, and applying a simplified annotation procedure. We used as few as 17 baseline features (Specia et al., 2009b) for QE training.

As the project progressed, post-edits became available, allowing us to validate our approach and further develop the bootstrapped system, using off-the-shelf PE distance (TER) as training labels. We added syntactic and web-scale Language Model (LM) features (Kozlova et al., 2016), (Andor, et al., 2016) to improve a second iteration of the QE system and trained on 80,000 PE distance labels to compare our results.

Finally, we roughly estimated the number of sentences needed for training a PE distance-based system that performs on par with a PE effort-based system.

2 Use case and related work

2.1 Use case

In Language Industry, Quality Estimation is used to filter out low-quality translations for post-editors, when they review Machine Translated texts (Specia et al., 2009b). This is important, because bad translations not only cause extra work (it is sometimes easier to translate from scratch (Specia, 2011)), they are also a source of frustration and negatively impact the image and acceptance of MT among translators (Wisniewski et al., 2013).

To alleviate these problems, we investigated the use of Quality Estimation for 9 language pairs (EN-DE, DE-EN, EN-FR, EN-RU, EN-ZH, EN-PT, EN-ES, EN-IT, EN-JP) and 3 domains (referred to as DOM1, DOM2 and DOM3). Since the MT engines were not cleared for use at the time the project began, no post-edits were available and a staged approach was required.

For production use, we are mainly interested in best practices (rather than in developing the best possible general-purpose QE system) and in deploying the system as quickly as possible with acceptable costs. This greatly differs from an academic setting, in which the exploration of Machine Learning algorithms and metrics, as well as the discovery of novel features are the main focus (see for example (Specia & Soricut, 2013)).

2.2 Related work

In industry, QE (also known as "Confidence Estimation (CE)" (Specia, 2011), (Blatz, et al., 2004) is most often used in sentence-based tasks, because all major translation environments use sentences as the basic units of work. For this reason, word-based (see for example (Blatz, et al., 2004), (Ueffing & Ney, 2005)) or document-based QE (see for example (Soricut & Echihabi, 2010)) were not considered, although they are useful in, respectively, the development of interactive MT systems, and document ranking for obtaining consistent high-quality output. The foundations of the work performed are described in (Callison-Burch et al., 2009), (Callison-Burch, et al., 2012) and (Specia et al., 2009b). We use their baseline system with the 17 features they describe.

3 Approach

Our approach differs in the way data collection is set up, and in the fact that we use PE effort judgments, although PE distance has been favored since the WMT 2013 campaign (Bojar, et al., 2013).

PE effort judgments were expressed according to the scores of (Callison-Burch, et al., 2012):

1. The MT output is incomprehensible, with little or no information transferred accurately. It cannot be edited, needs to be translated from scratch.
2. About 50-70% of the MT output needs to be edited. It requires a significant editing effort in order to reach publishable level.
3. About 25-50% of the MT output needs to be edited. It contains different errors and mistranslations that need to be corrected.
4. About 10-25% of the MT output needs to be edited. It is generally clear and intelligible.
5. The MT output is perfectly clear and intelligible. It is not necessarily a perfect translation but requires little or no editing.

The collection procedure outlined in WMT 2009 (Callison-Burch et al., 2009) was simplified as follows:

- By lack of post-edit data, neither high-quality targeted or hTER-optimized (Snover et al., 2006) post-edits were presented during annotation.
- No reference translation was presented – only the source sentence and MT output were displayed during annotation. Initial experiments showed that scores were assigned in too narrow a band when reference translations were provided. This potentially hurts QE performance, so we decided not to show them.
- We did not measure intra-annotator agreement, since we were dealing with professional translators, who are expected to perform similar tasks on a regular basis. Note that we intend not to discard any data.
- The obtained data was weighted according to the scheme in (Callison-Burch, et al., 2012): more weight was given to judges with higher standard deviation from their own mean score to obtain a more even spread in the range [1, 5].

We used the following metrics to evaluate our data sets and QE systems:

- Fleiss' coefficient (Fleiss, 1971), a generalization of Cohen's kappa to multi-raters (Wisniewski et al., 2013) to measure the degree of agreement between annotators.
- Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), standard metrics for regression, quantifying the amount by which the estimator differs from the true score (Specia et al., 2009a) (Wisniewski et al., 2013)
- Pearson's correlation, to express the linear correlation between predicted PE effort/PE distance and reference PE effort/PE distance.
- TER (Snover et al., 2006), to calculate the number of edits required to change a hypothesis translation into a reference translation.

Furthermore, we use our own proprietary software for feature extraction (based on (Eckart de Castilho & Gurevych, 2014)), and a LIBSVM epsilon-SVR with a Radial Basis Function Kernel, based on (Bethard et al., 2014).

Subsequent development of web-scale LM features is based on (Kozlova et al., 2016), the use of syntactic features is based on (Kozlova et al., 2016) and (Andor, et al., 2016).

4 Development of the baselines

4.1 MT Systems

The Machine Translation systems for which we develop QE, are based on Moses SMT (Koehn, et al., 2007), and on the work of (Neubig, 2013), and

DOMAIN	DE-EN MAE/MRSE		EN-DE MAE/MRSE		EN-ZH MAE/MRSE		EN-ES MAE/MRSE		EN-PT MAE/MRSE		EN-FR MAE/MRSE		EN-IT MAE/MRSE		ALL MAE/MRSE	
DOM1	0.65	0.88	***0.68***	0.88	-		-		-		-		-		0.73	0.97
DOM2	***0.54***	0.86	0.94	1.16	0.79	1.06	0.63	0.98	0.77	0.99	***0.54***	0.76	0.62	0.87	0.76	1.03
DOM3	-		0.80	1.05	***0.68***	0.95	*0.54*	0.85	0.86	1.10	0.63	0.95	-		0.79	1.03
LANG	0.63	0.90	0.80	1.03	0.70	0.97	***0.52***	0.83	***0.76***	1.02	0.55	0.80	0.62	0.87	0.77	1.04
BULK	0.77								1.04							

Table 3: QE test results

(Bisazza et al., 2011). The systems use extensive normalization, segmentation and classification routines, as well as some syntactic features. Since the focus is on QE, we will not go into further detail, but we list the data set sizes (number of unique sentence pairs) to give a general idea of the MT systems' potential output quality (see Table 1).

The domains consist of software-related materials, written in three distinctive styles. We will refer to them as DOM1, DOM2 and DOM3. DOM1 consists of solution descriptions, written by development and/or support staff, DOM2 relates to published documentation, DOM3 is intended for software training.

DOMAIN	DOM1	DOM2	DOM3
DE-EN	2,613,489	22,375,900	-
EN-DE	2,971,501	13,838,326	1,154,653
EN-ZH	-	2,557,042	439,980
EN-ES	-	3,456,275	366,423
EN-PT	-	2,942,499	298,687
EN-FR	-	4,944,361	343,352
EN-RU	-	2,108,723	455,203
EN-IT	-	3,198,050	-
EN-JP	878,036	4,915,823	533,053

Table 1: training set sizes MT systems

4.2 Data collection

The number of segments selected for each language pair is listed in Table 2. For DOM1 we only have 3 data sets and MT systems, but it is the only domain for which Post-Edits were available at the time of writing (see validation in section 5).

For each cell in the table, annotations from 3 translators were collected. Average inter-annotator agreement was at a level of 0.44 (Fleiss' coefficient) and can be considered *fair* according to (Landis & Koch, 1977).

DOMAIN	DOM1	DOM2	DOM3	TOTAL
DE-EN	800	800	-	1,600
EN-DE	800	800	800	2,400
EN-ZH	-	800	800	1,600
EN-ES	-	800	800	1,600
EN-PT	-	800	800	1,600
EN-FR	-	800	800	1,600
EN-RU	-	800	800	1,600
EN-IT	-	800	-	800
EN-JP	800	800	800	2,400

Table 2: training set sizes (PE Effort) QE systems

4.3 Results

The MAE and MRSE for the resulting systems are listed in Table 3. We tried several combinations of the data to find the optimum set of models:

- for each data set, ***language + domain-specific*** models were trained (listed in the white columns)
- ***language-specific*** models (LANG row) were trained by combining all data available for each language pair.
- language agnostic ***domain-specific*** models were trained by aggregating all data for each domain separately (ALL column in grey).
- finally, a language-agnostic **BULK** model (BULK row in grey), with all available data was trained.

The **BULK** model and the ***domain-specific*** models perform roughly on par, but in almost all cases, they are outperformed by the ***language-specific*** and ***language + domain-specific*** models. Which is what we expected, but we wanted to know whether it would be operationally feasible to train one single model or one model per domain.

In terms of performance, it is not clear which strategy, ***language-specific*** or ***language + domain-specific***, to select. From a systems management perspective though, having one ***language-specific*** system for each language pair reduces deployment complexity immensely, with only a minor decrease in performance as trade-off (except for the EN-DE language pair).

5 Validation of the approach

As stated in section 1, we fell back to the 2009 WMT protocol (Callison-Burch et al., 2009) by lack of PE data. We surmised that a prohibitive number of Post-Edits would be required to obtain acceptable QE performance, so only 800 segments (per domain and language-pair) were sent out for PE effort judgment (to 3 annotators) to remain within budget. If we assume – for the sake of simplicity – that annotating a sentence with a PE effort judgment and post-editing a sentence are

equally expensive, then we expect our bootstrapped ***language + domain-specific*** systems to outperform QE systems trained on three times as many PE distance labels (2,400 data points).

Figure 1 summarizes and extrapolates the number of data points it takes to obtain comparable correlations. The graphs clearly indicate that more than triple the data is required to get comparable QE performance. For EN-DE, we were able to obtain around 80.000 post-edits. Even with this relatively large data set, the baseline PE distance-based QE system does not achieve the quality we get from a PE effort-based system.

This corroborates our intuition that – starting with almost no data – it pays off to consider PE effort-based solutions when developing a baseline. Obviously, it would go too far to state that using PE effort should be the preferred, authoritative (Callison-Burch et al., 2009) approach, because there are too many intrinsic shortcomings to adopt it as a best practice. For example, the Pearson correlation we used to compare PE effort-based and PE-distance based QE, expresses the extent to which a predicted entity (PE effort or PE distance) has a linear relationship with *some* hidden variable. For all we know, this hidden variable may be *sentence length*, instead of Post-Edit quality. There is also the issue of *subjectivity* at the annotator side. PE distance eliminates subjectivity, and can thus be expected to yield more consistent results. We believe however, that the use of professional translators filtered out a lot of the noise that can be observed in the WMT campaigns.

Conversely, the extrapolation gives us an idea about how many sentence pairs are needed to build a system that performs on par with PE effort-based QE, using Post-Edits exclusively. This opens possibilities when training MT and QE systems in a data-rich (MT training data > 1M sentence pairs) environment. It would be interesting to investigate whether an optimum split can be achieved to divide the data into a larger part that is used to train MT systems with, and a smaller part that can be used to generate *pseudo* post-edits (the PE distance between reference and MT-generated hypothesis would be measured). The aim would be to maximize QE quality while minimizing MT quality loss. With the available data set, the use of *real* Post-Edits versus *pseudo* Post-Edits could be compared to validate such approach.

	System	Mae	Pearson Correlation
En-It	Baseline	0.628+/-0.029	0.460+/-0.050
	+OOVs+WLM	0.631+/-0.026	0.463+/-0.027
En-Fr	Baseline	0.543+/-0.024	0.367+/-0.028
	+OOVs+WLM	0.549+/-0.017	0.354+/-0.009
En-Pt	Baseline	0.766+/-0.012	0.416+/-0.010
	+OOVs+WLM	0.763+/-0.010	0.422+/-0.021
De-En	Baseline	0.597+/-0.014	0.486+/-0.015
	+OOVs+WLM	0.597+/-0.012	0.484+/-0.032
En-Ru	Baseline	0.624+/-0.015	0.335+/-0.030
	+OOVs+WLM	0.624+/-0.006	0.336+/-0.018
En-Es	Baseline	0.525+/-0.018	0.293+/-0.022
	+OOVs+WLM	0.522+/-0.018	0.304+/-0.012
En-Jp	Baseline	0.699+/-0.012	0.526+/-0.013
	+OOVs+WLM	0.719+/-0.012	0.499+/-0.014
En-De	Baseline	0.800+/-0.013	0.514+/-0.019
	+OOVs+WLM	0.794+/-0.009	0.520+/-0.006
En-Zh	Baseline	0.655+/-0.008	0.586+/-0.013
	+OOVs+WLM	0.657+/-0.007	0.586+/-0.003
Avg.	Baseline	0.649+/-0.006	0.442+/-0.008
	+OOVs+WLM	0.651+/-0.005	0.441+/-0.006

Table 4: comparison with and without OOVs and Web-scale LM

6 Additional features

Having obtained acceptable performance with a basic feature set, we added three features/feature sets to improve our models: technical OOVs, web-scale Language Models (WLMs) and SyntaxNet features.

6.1 Technical OOVs

When applying QE to real-life data, we expect the presence of technical OOVs (Fishel & Sennrich, 2014) to hurt performance for the following reasons: (1) usually, technical OOVs are not modelled in the MT system's translation and language model, instead they are normalized or treated as OOVs to be copied verbatim into the target. If this

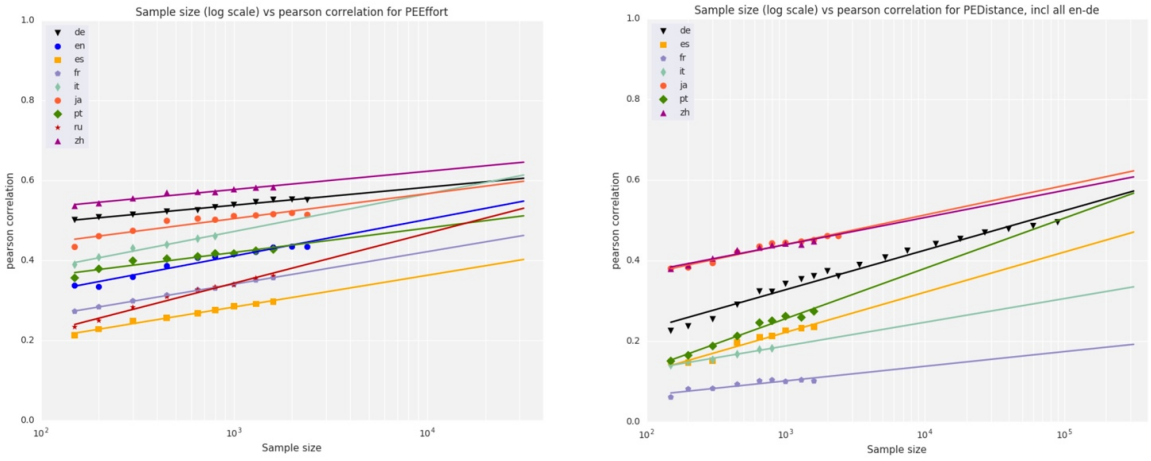

Figure 1: extrapolation of required PE distance labels for comparable performance

behaviour is not compensated for by the QE system, sentences with technical OOVs will unrightfully receive a penalty at lookup time; (2) in addition, technical OOVs, require a simple copy operation (if not resolved by the MT system), which makes the task of sentences containing OOVs easier, instead of more difficult

We use a custom-made classifier learnt from manually annotated data, and pre-processed with manually constructed rules (Kluegl et al., 2016), to annotate the training data.

6.2 Web-scale Language Models

We further experimented with Web-scale Language Models, as described in (Kozlova et al., 2016). We use public data (mostly Wikipedia) and collect around 48M sentences for English. The obtained gains are rather poor, probably because our language models are already quite big, and the extra out-of-domain data only adds little information.

6.3 SyntaxNet features

As a final experiment, we parsed our data with SyntaxNet (Andor, et al., 2016) and followed the approach outlined by (Kozlova et al., 2016). We use their tree-based features, as well as their features derived from Part-Of-Speech (POS) tags and dependency roles. Experiments were run on the EN-DE PE distance data set, because it was the only data set we had available at that time.

Our final results are listed in Table 5. The quality jump obtained (7,000 vs. 70,000), and the increasing difference between baseline (technical OOVs included for source and target) and best system, indicate that – in the long run – PE distance based QE remains worthwhile pursuing.

SAMPLE SIZE	FEATURES SET	#	MAE	PEARSON CORRELATION
700	BASELINE	19	0.269 +/ 0.003	0.258 +/- 0.017
	+ SYNTAX	43	0.264 +/ 0.001	0.318 +/- 0.005
	+ SYNTAX + WLM	45	0.267 +/- 0.002	0.309 +/- 0.012
7,000	BASELINE	19	0.241 +/ -0.001	0.432 +/- 0.005
	+ SYNTAX	43	0.237 +/- 0.001	0.459 +/- 0.002
	+ SYNTAX + WLM	45	0.236 +/- 0.001	0.460 +/- 0.004
70,000	BASELINE	19	0.229 +/- 0.001	0.504 +/- 0.002
	+ SYNTAX	43	0.219 +/- 0.001	0.548 +/- 0.002
	+ SYNTAX +WLM	45	0.217 +/- 0.001	0.556 +/- 0.002

Table 5: final results on the EN-DE PE distance data set

7 Discussion and future work

We have described the development of QE systems with no access to post-edit data. While mainly building on the work previously done in the QE field, our contribution consists of the development of *a method to quickly build QE systems with minimal resources and a simplified annotation scheme*. We observed that using around 100k PE distance labels can produce a QE system that correlates equally strong with PE quality as a PE effort-based system trained on 800 sentence pairs. This is valuable information, as it allows for budget planning and opens opportunities to use *pseudo* Post-Edits instead of real Post-Edits.

In the future, we plan to investigate the use of such *pseudo* Post-Edits and describe a method to obtain an optimum trade-off between MT quality and PE quality when operating in data-rich environments. We will also further develop the syntax-based features, using the +40 parsers that are made available through the SyntaxNet project.

Acknowledgements

The authors wish to thank the reviewers for their helpful suggestions.

References

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., . . . Collins, M. (2016). Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*.

Bethard, S., Ogren, P., & Becker, L. (2014). ClearTK 2.0: Design Patterns for Machine Learning in UIMA. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 3289-3293). Reykjavik, Iceland: European Language Resources Association (ELRA).

Bisazza, A., Ruiz, N., & Federico, M. (2011). Fill-up versus interpolation methods for phrase-based SMT adaptation. *IWSLT*, (pp. 136-143).

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., . . . Ueffing, N. (2004). Confidence estimation for machine translation. *Proceedings of the 20th international conference on Computational Linguistics* (pp. 315-321). Association for Computational Linguistics.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., . . . Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. *Proceedings of the Eighth Workshop on Statistical Machine Translation* (pp. 1-44). Sofia, Bulgaria: Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2007). (Meta-) evaluation of machine translation. *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 136-158). Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., & Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In A. f. Linguistics (Ed.), *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, (pp. 1-28). Athens, Greece.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., & Specia, L. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. *Proceedings of the Seventh Workshop on Statistical Machine Translation* (pp. 10-51). Stroudsburg, PA, USA: Association for Computational Linguistics.

Eckart de Castilho, R., & Gurevych, I. (2014, August). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT* (pp. 1-11). Dublin, Ireland: Association for Computational Linguistics and Dublin City University.

Fishel, M., & Sennrich, R. (2014). Handling Technical OOVs in SMT. *The Seventeenth Annual Conference of the European Association for Machine Translation (EAMT2014)*, (pp. 159-162).

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378.

Kluegl, P., Toepfer, M., Beck, P.-D., Fette, G., & Puppe, F. (2016). UIMA Ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering, 22*(01), 1-40.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., . . . Zens, R. (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177-180.

Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations. *Proceedings of the Seventh Workshop on Statistical Machine Translation* (pp. 181-190). Association for Computational Linguistics.

Kozlova, A., Shmatova, M., & Frolov, A. (2016). YSDA Participation in the WMT'16 Quality Estimation Shared Task. *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics. 2*, pp. 793-799. Berlin, Germany: Association for Computational Linguistics.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Neubig, G. (2013, August). Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers. *Proceedings of the ACL Demonstration Track*.

Neubig, G., Watanabe, T., & Mori, S. (2012). Inducing a discriminative parser to optimize machine translation reordering. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 843-853). Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of the Association for Machine Translation in the Americas*, *200*, pp. 223-231.

Soricut, R., & Echihabi, A. (2010). Trustrank: Inducing trust in automatic translations via ranking. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 612-621). Association for Computational Linguistics.

Specia, L. (2011). Exploiting objective annotations for measuring translation post-editing effort. *Proceedings of the 15th Conference of the European Association for Machine Translation*, (pp. 73-80).

Specia, L., & Farzindar, A. (2010). Estimating machine translation post-editing effort with HTER. *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10)*, (pp. 33-41).

Specia, L., & Soricut, R. (2013). Quality estimation for machine translation: preface. *Machine Translation, 27*(3-4), 167-170.

Specia, L., Saunders, C., Turchi, M., Wang, Z., & Shawe-Taylor, J. (2009a). Improving the confidence of machine translation quality estimates. *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*, (pp. 136-143).

Specia, L., Shah, K., De Souza, J. G., & Cohn, T. (2013). QuEst - A translation quality estimation framework. *ACL (Conference System Demonstrations)* (pp. 79-84). ACL.

Specia, L., Turchi, M., Cancedda, N., Dymetman, M., & Cristianini, N. (2009b). Estimating the sentence-level quality of machine translation systems. *13th Conference of the European Association for Machine Translation*, (pp. 28-37).

Ueffing, N., & Ney, H. (2005). Application of word-level confidence measures in interactive statistical machine translation. *Proceedings of EAMT*, (pp. 262-270).

Wisniewski, G., Singh, A. K., & Yvon, F. (2013). Quality estimation for machine translation: Some lessons learned. *Machine translation, 27*(3-4), 213-238.