# Supervised and Unsupervised Minimalist Quality Estimators: Vicomtech's Participation in the WMT 2018 Quality Estimation Task

**Thierry Etchegoyhen** and **Eva Martínez Garcia** and **Andoni Azpeitia**
Vicomtech
Mikeletegi Pasalekua, 57
Donostia / San Sebastián, Gipuzkoa, Spain
{tetchegoyhen, emartinez, aazpeitia}@vicomtech.org

## Abstract

We describe Vicomtech's participation in the WMT 2018 shared task on quality estimation, for which we submitted minimalist quality estimators. The core of our approach is based on two simple features: lexical translation overlaps and language model cross-entropy scores. These features are exploited in two system variants: uMQE is an unsupervised system, where the final quality score is obtained by averaging individual feature scores; sMQE is a supervised variant, where the final score is estimated by a Support Vector Regressor trained on the available annotated datasets. The main goal of our minimalist approach to quality estimation is to provide reliable estimators that require minimal deployment effort, few resources, and, in the case of uMQE, do not depend on costly data annotation or post-editing. Our approach was applied to all language pairs in sentence quality estimation, obtaining competitive results across the board.

## 1 Introduction

Quality Estimation (QE) refers to the task of estimating the quality of machine translation output without access to reference translations (Blatz et al., 2004), which are not always available for a given domain or language pair, and are costly to produce.

Typical approaches are based on supervised machine learning models using a large array of features, as exemplified by the standard QUEST baseline (Specia et al., 2013), whose base version employs 17 features that include n-gram language model perplexity scores, lexical translation probabilities, number of source tokens and average number of translations per source word, among others. In recent years, QE models based on neural network approaches have significantly improved the state of the art, as shown for instance by the results obtained in the WMT 2016 and WMT 2017 shared tasks (Kim and Lee, 2016; Kim et al., 2017; Martins et al., 2017).

Despite recent progress, the vast number of potential domains and language pairs is a challenging aspect for a practical use of quality estimation systems. First, most approaches to QE rely on annotated data, typically based on human post-editing, which are costly to produce. Additionally, the best performing approaches based on neural networks (e.g., Kim et al., 2017) require large volumes of parallel training corpora, a resource which is only available for a small number of language pairs nowadays.

To tackle these challenges, we designed a minimalist approach to quality estimation, to which we will refer as MQE, based on two features: a lexical translation overlap measure to model translation accuracy[1] and a measure based on cross-entropy scores according to a target language model. No external tools or large computational resources are needed in this approach, which can be used in the two variants described below.

uMQE is an unsupervised variant, where the final quality score is obtained by averaging individual feature scores. The system was designed to provide reliable estimators in the numerous use cases where no training data are available to train supervised QE models. To our knowledge, little attention has been paid to this type of approaches, with two main published approaches: Moreau and Vogel (2012) estimate the quality of machine-translated output against external sets of n-grams and evaluate several variants of n-gram similarity, whereas Popovic (2012) proposes an unsupervised method based on the arithmetic combination of scores provided by language models and IBM1 models, trained on morphemes as well as part-of-speech tags. On the WMT 2012 datasets, neither

[1] Also referred to as *adequacy*.



*Proceedings of the Third Conference on Machine Translation (WMT)*, Volume 2: Shared Task Papers, pages 795–800

https://doi.org/10.18653/v1/W18-6462

approach performed better than the QUEST baseline. In this paper, we show that our own unsupervised approach can outperform the supervised baselines, without the use of additional resources such as part-of-speech taggers or morphological analysers.

sMQE is a supervised variant, where the final score is estimated by a Support Vector Regressor trained on the available machine translation output annotated with HTER scores. The goal of this approach is to enable a fast deployment of supervised quality estimators that outperform other supervised approaches with more complex setups, such as the QUEST baselines with 17 features, while using minimal resources. Contrary to uMQE, for which only rank correlation is meaningful, the supervised variant can be evaluated on both ranking and scoring tasks.

The paper is organised as follows: Section 2 describes the core MQE approach and the computation of the supervised and unsupervised variants; Section 3 describes the experimental setup for the WMT 2018 shared task on sentence quality estimation; Section 4 presents our results on the test sets in all four language pairs and domains; finally, Section 5 draws conclusions from this work.

## 2 MQE

Minimally, quality estimation involves determining the accuracy (or adequacy) of a translation, i.e. how much of the source information is represented in the translation, and its fluency, i.e. the correctness of the generated sentence as a target language sequence. MQE directly models these two aspects, to the exclusion of any other property of the source and target sentence pairs. We describe our measures of accuracy and fluency in turn in the next sections.

### 2.1 Accuracy

To measure accuracy, we adapted the approach in (Etchegoyhen and Azpeitia, 2016), which has proved highly successful in identifying parallel sentences in large sets of comparable corpora (Azpeitia et al., 2017, 2018). Their method is based on Jaccard similarity (Jaccard, 1901) over lexical sets, with additional set expansion operations to address named entities and morphological variation. We describe their core methodology below and our adaptations for the quality estimation task.

Let $s_i$ and $s_j$ be two tokenised and truecased sentences in languages $l_1$ and $l_2$, respectively, $S_i$ and $S_j$ the multisets[2] of tokens in $s_i$ and $s_j$, respectively, $T_{ij}$ the multiset of lexical translations into $l_2$ for all tokens in $S_i$, and $T_{ji}$ the multiset of lexical translations into $l_1$ for all tokens in $S_j$.

Lexical translations are computed from sentences $s_i$ and $s_j$ by retaining the $k$-best translations for each word, as determined by the ranking obtained from the translation probabilities given by symmetrised IBM2 word alignment models (Brown et al., 1993).[3] The multisets $T_{ij}$ and $T_{ji}$ that comprise these $k$-best lexical translations are then expanded by means of the following operations:[4]

1. For each element $x$ in the set difference $T'_{ij} = T_{ij} - S_j$ (respectively $T'_{ji} = T_{ji} - S_i$), and each element $y$ in $S_j$ (respectively $S_i$), if $x$ and $y$ share a common prefix of more than $n$ characters, the prefix is added to both $T_{ij}$ and $S_j$ (respectively $T_{ji}$ and $S_i$). Longest common prefix matching is meant to capture morphological variation via minimal computation.

2. Numbers and capitalised truecased tokens not found in the lexical translation tables are added to the expanded translation multisets $T_{ij}$ and $T_{ji}$. This operation addresses named entities, which are likely to be missing from translation tables trained on different domains.

3. The NULL token is added to the source and target token multisets, in order to address words that have covert translations, as indicated by the presence of the NULL element among their $k$-best translation options.

With source and target sets as defined above, we compute translation accuracy between sentence $s_i$ and translation $s_j$ as in Equation 1:

[2] We employ multisets instead of sets as in the original approach, to account for multiple token occurrences, as the quality estimation task is more likely to be sensitive to missing occurrences than the alignment task. Multiset intersection and union are based on positive minimums and maximums, respectively.

[3] The actual probabilities are not used beyond determining the ranking, as in the original approach. We depart from their implementation by using IBM2 models instead of IBM4, a change motivated by the similar results we obtained with both types of models and the faster training of the former.

[4] The first two are based on the original approach, while the third was added by us for the experiments reported here.

$$acc(s_i, s_j) = \frac{1}{2}\left(\frac{|T_{ij} \cap S_j|}{|S_j|} + \frac{|T_{ji} \cap S_i|}{|S_i|}\right) \quad (1)$$

Accuracy is thus defined as the mean of the overlap similarity coefficients obtained between sentence token sets and expanded lexical translation sets in both directions.[5] Apart from the use of multisets and the introduction of the NULL element, the main change to the original metric is using overlap instead of Jaccard similarity, as the former provided better results in preliminary experiments.

Although originally meant to identify parallel sentences in comparable corpora, this simple metric applies naturally to any task involving lexical translations and provides an efficient method to model accuracy.

### 2.2 Fluency

The standard approach to measuring the fluency of word sequences in a given language is by means of language models. Although n-gram modelling has been the dominant approach in the last two decades, continuous space language models have become a new standard and have been notably used for the quality estimation task, providing improvements in supervised feature-based frameworks (Shah et al., 2015b). For the experiments presented here, we nonetheless used n-gram language modelling as a first approach, as they provided the best results overall in preliminary experiments and require comparatively fewer computational resources to be trained.

As a measure of fluency, we take the inverse of the per word cross-entropy for each machine-translated sentence. The fluency score is thus computed according to Equation 2, where $P(w_i)$ is short for $P(w_i|w_{i-(k-1)}, \ldots, w_{i-1})$, i.e. the conditional probability of the $i$-th word given its $k$ preceding words in sentence $s_j$ of length $n$.

$$\mathit{flc}(s_j) = \frac{1}{-\frac{1}{n}\sum_{i=1}^{n} logP(w_i)} \quad (2)$$

Thus, the higher the cross-entropy, the lower the fluency score. Although simple, measures computed via n-gram language models, such as cross-entropy or the monotonically-related perplexity, have been shown to be reliable indicators of translation quality estimation (Shah et al., 2015a).

### 2.3 MQE Variants

For the unsupervised uMQE variant, we assume that task-related annotated data are not available to optimise feature weighting,[6] and thus simply take the arithmetic mean of the two scores as our final quality estimation score. Since the two scores are not in similar ranges, we perform min/max feature rescaling on all scores prior to combining the features. The final quality estimation score for a source $s_i$ and translation $s_j$ is computed as in Equation 3, with rescaled features $acc^r$ and $flc^r$.

$$\mathit{uMQE}(s_i, s_j) = \frac{acc^r(s_i, s_j) + flc^r(s_j)}{2} \quad (3)$$

For the supervised variant, sMQE, we used the annotated datasets provided for the WMT 2018 QE task and trained a Support Vector Regressor (SVR) with a Radial Basis Function kernel on the two features, using the default parameters provided by the scikit-learn toolkit[7] (C=1.0, $\epsilon$=0.1, and $\gamma$=0.5 for 2 features):

$$\mathit{sMQE}(s_i, s_j) = \mathit{SVR}([acc(s_i, s_j), flc(s_j)]) \quad (4)$$

## 3 Experimental Setup

We submitted results from our two system variants in all language pairs for sentence-level QE, using the same models for both variants in each case. To train the IBM2 and language models, we selected corpora available for the WMT shared tasks for each specific domain and language pair. For English-German, in the IT domain, we used the training data from the WMT 2016 IT translation task, the WMT 2017 QE task and the WMT 2018 PE task; given the low amounts of data in each individual corpus, we also merged the data from the technical manuals of *OpenOffice* and *KDE4* available in the OPUS repository (Tiedemann, 2012). For German-English, in the biomedical domain, we used the UFAL medical corpus[8], combined with the training data from the WMT 2018 QE

[5] Note that the denominator in a set-based overlap measure is the smallest of the two sets being compared, which in our case is always the token set.

[6] Such datasets were available for the WMT 2018 shared task, but we opted to ignore them in order to test the uMQE variant under its intended unsupervised conditions of use.

[7] http://scikit-learn.org/

[8] https://ufal.mff.cuni.cz/ufal_medical_corpus

| LANG | DOMAIN | MT | SYSTEM | SPEARMAN | $RK^{\rho}$ | PEARSON | $RK^{\tau}$ | MAE | RMSE |
|---|---|---|---|---|---|---|---|---|---|
| EN-DE | IT | SMT | UMQE | 0.3787 | 12/15 | _ | _ | _ | _ |
| EN-DE | IT | SMT | UMQE* | 0.4042 | 7/15 | _ | _ | _ | _ |
| EN-DE | IT | SMT | SMQE | 0.3993 | 7/15 | 0.3969 | 9/14 | 0.1855 | 0.2248 |
| EN-DE | IT | NMT | UMQE | 0.3999 | 10/13 | _ | _ | _ | _ |
| EN-DE | IT | NMT | UMQE* | 0.4542 | 6/13 | _ | _ | _ | _ |
| EN-DE | IT | NMT | SMQE | 0.4439 | 6/13 | 0.3716 | 9/12 | 0.2063 | 0.2421 |
| DE-EN | BIOMED | SMT | UMQE | 0.5694 | 5/11 | _ | _ | _ | _ |
| DE-EN | BIOMED | SMT | SMQE | 0.6003 | 4/11 | 0.6521 | 4/10 | 0.1182 | 0.1547 |
| EN-LV | BIOMED | SMT | UMQE | 0.3979 | 3/8 | _ | _ | _ | _ |
| EN-LV | BIOMED | SMT | SMQE | 0.4061 | 2/8 | 0.4612 | 2/7 | 0.1318 | 0.1767 |
| EN-LV | BIOMED | NMT | UMQE | 0.5403 | 3/7 | _ | _ | _ | _ |
| EN-LV | BIOMED | NMT | SMQE | 0.5686 | 2/7 | 0.5787 | 2/6 | 0.1461 | 0.1938 |
| EN-CS | IT | SMT | UMQE | 0.4196 | 6/9 | _ | _ | _ | _ |
| EN-CS | IT | SMT | SMQE | 0.4219 | 5/9 | 0.3904 | 7/8 | 0.1638 | 0.2122 |

Table 1: Results on the WMT 2018 test sets

task. For English-Latvian, also in the biomedical domain, we used the available EMEA corpus along with the training data from the WMT 2018 QE task and the additional data provided for this language pair in this year's QE task. Finally, for English-Czech in the IT domain, we used the *train-techdoc* section of the CzEng17 dataset available for the WMT 2018 translation task, along with the QE training data and the additional data provided for the WMT 2018 QE task.

Sentences were tokenised and truecased with the scripts available in the Moses toolkit (Koehn et al., 2007), with truecasing models trained on the data described above. For English-Czech, we experimented with BPE segmentation (Sennrich et al., 2016) to overcome data sparseness issues, training BPE models with a maximum of 30.000 merge operations and segmenting all corpora accordingly for this language pair.

All IBM2 models were trained with the FASTALIGN toolkit (Dyer et al., 2013), and all language models are of order 5 trained with the KENLM toolkit (Heafield, 2011) on the target language data. For the accuracy metric, minimal prefix length was set to 4 and $k$-best translation lists limited to 4 candidates.

## 4 Results

The results on the WMT 2018 test sets are shown in Table 1.[9] Overall the results were satisfactory for both variants of such a simple minimalist approach. For English-Latvian for instance, sMQE and uMQE ranked in second and third place, respectively; for German-English, the two variants ranked fourth and fifth, respectively. Our worst results were obtained for English-Czech and English-German, although for the latter our system still ranked in the top half among competing systems on the ranking task, and, except for the scoring task in EN-CS, both variants outperformed the baselines across the board. The relatively worse results obtained for these two language pairs can be tied to data sparseness issues affecting our simple fluency feature based on n-gram cross-entropy.

The results obtained by uMQE were overall slightly lower than those obtained by the supervised sMQE variant, although the small number of features available to train the SVR for the latter was not expected to lead to major improvements. Our unsupervised approach gave satisfactory results, performing significantly better in most cases than the supervised baseline with 17 features. We view this as an important result, considering the vast number of domains and language pairs where no training data are available to opt for a supervised approach.

Even in cases where task-related data exist, the amount of available parallel corpora in a given language pair might not be sufficient to train sophisticated neural quality estimators. In such cases, the sMQE variant can also provide a reliable alternative to perform quality estimation under minimal resources.

The approach is also fairly simple to implement and deploy, and does not require external tagging or parsing tools which may not be available for

[9] In the table, $RK^{\rho}$ and $RK^{\tau}$ indicate the ranking of the system among all participants in terms of Spearman and Pearson correlation, respectively. Note that the official uMQE results for English-German are based on erroneous submissions and we submitted the correct version after the deadline via CODALAB to obtain the expected scores. The correct version, using the same models as for sMQE, is denoted by uMQE* and we refer to the results of this submission in the discussion relative to this language pair.

many languages. It is thus a highly portable alternative which may be the simplest and most efficient option in a significant number of scenarios, with results that outperform the standard supervised baseline across the board.

## 5 Conclusions

We have described our participation in the WMT 2018 shared task on quality estimation, which included both supervised and unsupervised variants of a minimalist approach to the task. Both variants are based on two simple measures of accuracy, computed from lexical translation overlap, and fluency, computed from inverse cross-entropy scores of an n-gram language model.

Our main goal was to evaluate systems that can be efficiently deployed for the large number of language pairs and domains where there are either no annotated data at all to train a supervised system, or insufficient amounts of parallel corpora to adequately train the currently best performing neural quality estimators. Additionally, our approach requires no external tools such as part-of-speech taggers or syntactic parsers, unlike other competing approaches, and is thus both simpler to deploy and readily available for languages where such tools are not available at all.

We view the obtained results as satisfactory, with both variants outperforming the supervised baselines overall and being placed among the five best systems in two of the four language pairs. In future work, we will evaluate the use of continuous space language models to address data sparseness issues in the two language pairs where more complex morphology limits the contribution of an n-gram-based fluency feature. We will also explore variants of the accuracy measure and evaluate in more details the aspects that can be better modelled under the proposed minimalist approach to quality estimation.

**Acknowledgements** This work was supported by the Department of Economic Development and Competitiveness of the Basque Government via project QUALES (KK-2017/00094). We wish to thank the anonymous WMT 2018 reviewers for their detailed and helpful reviews of this work. Opinions, errors and omissions are our own.

## References

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. 2017. Weighted Set-Theoretic Alignment of Comparable Sentences. In *Proceedings of the Tenth Workshop on Building and Using Comparable Corpora*, pages 41–45. Association for Computational Linguistics.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. 2018. Extracting Parallel Sentences from Comparable Corpora with STACC Variants. In *Proceedings of the Eleventh Workshop on Building and Using Comparable Corpora*, pages 48–52. European Language Resources Association (ELRA).

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence Estimation for Machine Translation. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 315–321. Association for Computational Linguistics.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational linguistics*, 19(2):263–311.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Thierry Etchegoyhen and Andoni Azpeitia. 2016. Set-Theoretic Alignment for Comparable Corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 2009–2018.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Paul Jaccard. 1901. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241 – 272.

Hyun Kim and Jong-Hyeok Lee. 2016. Recurrent Neural Network based Translation Quality Estimation. In *Proceedings of the First Conference on Machine Translation*, pages 787–792. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared*

*Task Papers*, pages 562–568. Association for Computational Linguistics.

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 *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180. Association for Computational Linguistics.

André F. T. Martins, Fabio Kepler, and Jose Monteiro. 2017. Unbabel's participation in the wmt17 translation quality estimation shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 569–574. Association for Computational Linguistics.

Erwan Moreau and Carl Vogel. 2012. Quality estimation: an experimental study using unsupervised similarity measures. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 120–126. Association for Computational Linguistics.

Maja Popovic. 2012. Morpheme- and pos-based IBM1 and language model scores for translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 133–137.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Kashif Shah, Trevor Cohn, and Lucia Specia. 2015a. A bayesian non-linear method for feature selection in machine translation quality estimation. *Machine Translation*, 29(2):101–125.

Kashif Shah, Raymond WM Ng, Fethi Bougares, and Lucia Specia. 2015b. Investigating continuous space language models for machine translation quality estimation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1073–1078.

Lucia Specia, Kashif Shah, Jose GC de Souza, Trevor Cohn, and Fondazione Bruno Kessler. 2013. QuEst–a translation quality estimation framework. In *Proceedings of the 51st ACL: System Demonstrations*, pages 79–84.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th Language Resources and Evaluation Conference*, pages 2214–2218.