

Automatic Annotation of Machine Translation Datasets with Binary Quality Judgements

Marco Turchi, Matteo Negri

FBK, Fondazione Bruno Kessler
38123 Povo, Trento, Italy
{turchi,negri}@fbk.eu

Abstract

The automatic estimation of machine translation (MT) output quality is an active research area due to its many potential applications (e.g. aiding human translation and post-editing, re-ranking MT hypotheses, MT system combination). Current approaches to the task rely on supervised learning methods for which high-quality labelled data is fundamental. In this framework, quality estimation (QE) has been mainly addressed as a regression problem where models trained on (*source, target*) sentence pairs annotated with continuous scores (in the [0-1] interval) are used to assign quality scores (in the same interval) to unseen data. Such definition of the problem assumes that continuous scores are informative and easily interpretable by different users. These assumptions, however, conflict with the subjectivity inherent to human translation and evaluation. On one side, the subjectivity of human judgements adds noise and biases to annotations based on scaled values. This problem reduces the usability of the resulting datasets, especially in application scenarios where a sharp distinction between “good” and “bad” translations is needed. On the other side, continuous scores are not always sufficient to decide whether a translation is actually acceptable or not. To overcome these issues, we present an automatic method for the annotation of (*source, target*) pairs with *binary* judgements that reflect an empirical, and easily interpretable notion of quality. The method is applied to annotate with binary judgements three QE datasets for different language combinations. The three datasets are combined in a single resource, called BinQE, which can be freely downloaded from <http://hlt.fbk.eu/technologies/binqe>.

Keywords: Machine Translation, Quality Estimation, Corpus Annotation.

1. Introduction

In the machine translation (MT) field, quality estimation (QE) is the task of determining the quality of an automatic translation given its source sentence (Specia et al., 2009; Soricut and Echihabi, 2010; Bach et al., 2011; Specia, 2011; Mehdad et al., 2012). Differently from standard MT evaluation methods that rely on metrics such as BLEU (Papineni et al., 2002), QE aims to return predictions for unseen translated sentences without relying on reference translations. This makes QE particularly suitable from an application-oriented perspective, since in many situations the quality of automatic translations has to be measured at run-time and without the support of external, manually created benchmarks.

Among the many potential applications, quality estimates are extremely useful in computer-assisted translation (CAT) where, for each segment of a source document, human translators are presented with suggestions obtained from a translation memory (TM) or an MT engine. In both cases, to enhance translators’ productivity, *useful* suggestions (whose correction requires less effort than re-translation from scratch) should be promoted, and the *useless* ones should be demoted or automatically filtered out.

While TM suggestions are typically accompanied by *fuzzy match* scores (indicating the amount of overlap between the source sentence and previously translated segments stored in the translation memory), MT outputs can be presented with different quality indicators that account for their reliability. Such indicators typically consist in *effort score* labels (indicating the expected amount of post-editing needed by a human to correct a translation into a publishable material), or *time* estimates (indicating the expected time in seconds needed for the correction). Besides the fact that these qual-

ity indicators are not comparable with fuzzy match scores,¹ the idea that translation quality can be represented with scaled values raises other issues related to their use and interpretation in the CAT framework.

The first problem is that quality judgements are subjective (Koponen, 2012; Turchi et al., 2014), as also evidenced by the low inter-annotator agreement on the available datasets (Callison-Burch et al., 2012). Since different annotators often produce different quality scores for the same (*source, target*) pair, the resulting datasets are usually affected by various levels of noise and bias in labels’ distribution. This issue complicates the task of learning reliable QE models and can drastically reduce their usability.

Another problem, also related to the subjectivity of human judgements, is that continuous quality scores are not easily interpretable. For instance, a *0.49 effort score* does not say much about the actual quality of a translation, nor about how different users will perceive it.

An intuitive solution to make QE judgements suitable for practical use is to approach the problem as a binary task. Our hypothesis is that, especially for some applications such as CAT technology, QE models trained on binary datasets would make possible to obtain quality judgements that are more informative and easily interpretable than continuous scores. To this aim, the existing datasets can be transformed into binary datasets by setting a threshold that discriminates between “good” and “bad” translations. Following such thresholding method, instances with an effort score (or time in seconds) above the threshold would be marked as *bad* (i.e. useless) while those below the threshold would be marked as *good* (i.e. useful). This strategy,

¹This problem is out of the scope of our investigation, which exclusively focuses on making QE judgements more suitable for practical use in binary tasks.

however, has to confront with the problem (still related to human subjectivity) of setting appropriate, objective and reliable thresholds. Indeed, any threshold used to map the labels of an existing dataset into the two classes would be arbitrary and not necessarily acceptable by different users. To cope with these issues, in (Turchi et al., 2013) we proposed an automatic method for the annotation of (*source*, *target*) sentence pairs with binary judgements that reflect an empirical, non subjective and easily interpretable notion of quality. Taking advantage of its effectiveness, in this paper we applied such method to re-annotate with binary judgements three existing QE datasets for different language pairs.

The remainder of the paper is structured as follows: Section 2. briefly explains our automatic annotation method. Sections 3. and 4. respectively describe the three re-annotated datasets and conclude the paper.

The three binary datasets annotated with our procedure combined in a single resource, called BinQE, which is distributed under a Creative Commons Attribution-NonCommercial-ShareAlike license and can be freely downloaded from <http://hlt.fbk.eu/technologies/binqe>.

To the best of our knowledge, these represent the first freely available QE corpora empirically (and reliably) annotated with binary quality labels.

2. Automatic annotation method

Current QE datasets usually contain *automatic translations* (hundreds to thousands) along with their *source sentences*, *reference translations*, *post-edited translations* and, in the training set, *quality labels*. The quality labels (either human judgements, post-editing time in seconds or distance scores, such as the HTER,² calculated between the target and its post-edited version) are typically used to train regression models, which are later used to label new unseen instances in a test set.

Moving to a binary classification scenario, our task is to re-annotate an existing dataset by assigning to each instance a binary label (*i.e.* -1/+1) that indicates the quality of the translation (respectively bad/good).

2.1. Approach

Our technique (Turchi et al., 2013), which does not involve subjective human judgements, is based on the observation of similarities and dissimilarities between an automatic translation (TGT), its post-edited version (PE) and the corresponding reference translation (RT). Such comparisons provide useful indications about the behaviour of a post-editor when correcting automatic translations and, in turn, about MT output quality.

Typically, the PE version of a good-quality TGT preserves some characteristics (*e.g.* lexical, structural) that indicate a moderate correction activity by the post editor. Conversely,

²The Human-targeted Translation Edit Rate (Snover et al., 2009) is the minimum edit distance between the machine translation and its manually post-edited version in the [0,1] interval. Possible editing operations include the insertion, deletion, and substitution of single words as well as shifts of word sequences.

in the PE version of a low-quality TGT, such characteristics are more difficult to observe, indicating an intense correction activity. At the two extremes, the PE of a perfect TGT preserves all its characteristics, while the PE of a useless TGT loses most of them. In the first case TGT and PE are identical, and their similarity is the highest possible (*i.e.* $sim(TGT, PE) = 1$). In the second case, TGT and PE show a degree of similarity close to that of TGT and a completely rewritten translation featuring different lexical choices and structure. This is where reference translations come into play: considering RT as a good example of rewritten sentence,³ for low-quality TGT we will have $sim(TGT, PE) \approx sim(TGT, RT)$.

In light of these considerations, we hypothesize that the automatic re-annotation of existing QE datasets can take advantage of a classifier that learns a similarity threshold T such that:

- A PE sentence with $sim(TGT, PE) \leq T$ will be considered as a rewritten translation (hence TGT is useless, and the corresponding *source-TGT* pair a negative example to be labelled as -1).
- A PE sentence with $sim(TGT, PE) > T$ will be considered as a real post-edition (hence TGT is useful for the post-editor, and the corresponding *source-TGT* pair a positive example to be labelled as +1).

Based on these hypotheses, our automatic re-annotation of existing QE datasets labelled with continuous values is carried out by:

1. Creating a training set Z of positive and negative examples (*i.e.* (TGT, *correct_translation*) pairs, where *correct_translation* is either a post-edition or a rewritten translation).
2. Designing a feature set capable to capture different aspects of the similarity between TGT and *correct_translation*.
3. Building a binary classifier using Z .
4. Using the classifier to re-label the (TGT, PE) pairs as instances of post-editions or rewritings.

As regards the first step, a training set Z is created from each dataset, which we aim to re-annotate by taking advantage of the available post-editions and reference translations. Concerning the features, our feature set has been designed to capture text similarity by measuring word/n-gram matches (*e.g.* calculating ROUGE scores), as well as the level of sparsity and density of the common words as a shallow indicator of structural similarity between TGT and *correct_translation*. Using these features an SVM classifier is trained and used to re-annotate Z with binary quality labels.

³Such assumption is supported by the fact that reference sentences are, by definition, free translations manually produced without any influence from the target.

2.2. Building the classifier

Training corpus. To build a classifier capable of labelling PE sentences as rewritten/post-edited material, for each dataset we first created a training corpus (Z) of positive and negative instances. For each tuple (SRC, TGT, PE, RT), one positive and one negative instance have been respectively obtained as the combination of (TGT, PE) and (TGT, RT).

Features. Crucial to our classification task, a number of features can be used to estimate sentence similarity. Differently from the binary QE task, where the possibility to catch common characteristics between two sentences is limited by language barriers, in our re-annotation task all the features are extracted by comparing two monolingual sentences (TGT and a *correct_translation*, either a PE or a RT).

Although the problem of measuring sentence similarity can be addressed in many ways, the solutions should not overlook the specificities of the task. In our case, for instance, the scarce importance of the semantic aspect (TGT, PE and RT typically show a high semantic similarity) makes features used for other tasks (*e.g.* based on distributional similarity) less effective than shallow features looking at the surface form of the input sentences.

Our problem presents some similarities with the plagiarism detection task, where subtle lexical and structural similarities have to be identified to spot suspicious plagiarized texts (Potthast et al., 2010). For this reason, part of our features (*e.g.* ROUGE scores) are inspired by research in such field (Chen et al., 2010), while others have been designed *ad-hoc*, based on the specific requirements of our task. The resulting feature set aims to capture text similarity by measuring word/n-gram matches, as well as the level of sparsity and density of the common words as a shallow indicator of structural similarity.

In total, from each (TGT, *correct_translation*) pair, the following 22 features are extracted:

- Human-targeted Translation Error Rate – HTER. The editing operations considered are: shift, insertion, substitution and deletion.
- Number of words in common.
- Number of words in common, normalized by TGT length and *correct_translation* length (2 features).
- Number of words in TGT and in the *correct_translation* (2 features).
- Size of the longest common subsequence.
- Size of the longest common subsequence, normalized by TGT length.
- Aligned word density: total number of aligned words,⁴ divided by the number of aligned blocks (more than 1 aligned word).

⁴Monolingual stem-to-stem exact matches between TGT and *correct_translation* are inferred by computing the HTER, as in (Blain et al., 2012).

- Unaligned word density: total number of unaligned words, divided by the number of unaligned blocks (more than 1 unaligned word).
- Normalized number of aligned blocks: total number of aligned blocks, divided by TGT length.
- Normalized number of unaligned blocks: total number of unaligned blocks, divided by TGT length.
- Normalized density difference: difference between aligned word density and unaligned word density, divided by TGT length.
- Modified Lesk score (Lesk, 1986): sum of the squares of the length of n-gram matches, normalized by the product of the sentence lengths.
- ROUGE-1/2/3/4: n-gram recall with $n=1, \dots, 4$ (4 features).⁵
- ROUGE-L: size of longest common subsequence, normalized by the *correct_translation* length.
- ROUGE-W: the ROUGE-L using different weights for consecutive matches of length L (default weight = 1.2).
- ROUGE-S: the ROUGE-L allowing for the presence of skip-bigrams (pairs of words, even not adjacent, in their sentence order).
- ROUGE-SU: the extension of ROUGE-S adding unigrams as counting unit.

To increase the possibility to identify text similarities, all sentences are tokenized, lowercased and stemmed using the Snowball algorithm (Porter, 2001).

Classifier. For each resulting corpus Z , an SVM classifier (Mammone et al., 2009) has been trained using the LIB-SVM toolbox (Chang and Lin, 2011). The selection of the kernel and the optimization of the parameters were carried out through grid search in 5-fold cross-validation.

3. Automatically re-annotated datasets

Our re-annotation procedure has been applied to label with binary judgements the following three datasets:

- **WMT13** - The QE dataset used for Task 1.1 at WMT 2013 (Bojar et al., 2013). This corpus consists of 2,754 English-Spanish news sentences (2,254 from the training and test sets of WMT 2012 and 500 from the test set of WMT 2013). Automatic translations were obtained from the phrase-based Moses SMT system (Koehn et al., 2007) trained on Europarl (Koehn, 2005) and News Commentaries corpora.⁶ Reference translations, post-edited translations, and HTER scores are also provided for each instance.

⁵All ROUGE scores, described in (Lin, 2004), have been calculated using the software available at <http://www.berouge.com>.

⁶<http://www.statmt.org/wmt11/translation-task.html>

- **AMT** - The French-English dataset described in (Potet et al., 2012). This corpus consists of 10,881 news sentences translated with a Moses-based SMT system (Potet et al., 2010), along with their reference translations and post-edited translations. Post-editions were collected using Amazon’s Mechanical Turk.⁷
- **CAT** - An English-Italian dataset collected within the MateCat EU-Project⁸. This corpus consists of 1,261 tuples from the legal domain. Source and reference sentences were extracted from four parallel documents of a European Parliament resolution published on the EUR-Lex platform.⁹ The source sentences were translated by the Moses toolkit trained on 1.5M parallel sentences extracted from the Acquis corpus (Steinberger et al., 2006). Post-editions were collected by a language service provider from professional translators operating with the MateCat CAT tool in real working conditions.

Table 1 provides basic information (language pairs, domain, total number of positive and negative instances) about each re-annotated dataset.

Dataset	Languages	Domain	#Instances	#Positive	#Negative
WMT	EN-SP	News	2,754	2,022	732
AMT	FR-EN	News	10,881	8,847	2,034
CAT	EN-IT	Legal	1,261	685	576

Table 1: QE datasets re-annotated with binary judgements.

The quality of our re-annotation has been extrinsically evaluated in (Turchi et al., 2013). To this aim, the performance of binary QE classifiers trained on data produced with our automatic method has been compared with the classification results obtained by models trained on arbitrary partitions of the original corpora. Such partitions were obtained by thresholding the labels of the original datasets (either HTER scores, human effort scores, or post-editing time values) in different ways. Evaluation results show that all the models trained on arbitrary partitions are significantly outperformed by those trained on our data-driven re-annotation. This holds not only for arbitrary partitions generating highly unbalanced distributions of positive and negative examples (*i.e.* *good* and *bad* translations),¹⁰ but also for those generating balanced distributions. This suggests that the quality of the separation is as important as the actual proportion of positive and negative instances and that, in our case, it is superior to the quality of annotation schemes based on arbitrary thresholding schemes.

⁷<http://www.mturk.com>

⁸<http://www.matecat.com/>

⁹<http://eur-lex.europa.eu/>

¹⁰For some thresholds, the corresponding partitions of the WMT dataset produce highly unbalanced datasets. For instance, with the 0.7 HTER threshold proposed by the WMT guidelines as a reasonable boundary between *useful* and *useless* translations, the proportion between positive and negative examples is 0.02%. QE classifiers trained on such unbalanced distributions typically perform majority voting, thus achieving poor performance results.

4. Conclusion

We discussed the application of a data-driven technique for the re-annotation with binary quality judgements of existing QE datasets labelled with continuous scores (*e.g.* HTER or post-editing time) or Likert values. Our method has been applied to re-annotate three existing datasets for different domains and language combinations. The evaluation of models trained on the resulting corpora, which was carried out in (Turchi et al., 2013), shows that our empirical labelling method produces more accurate annotations than those obtained by partition methods based on arbitrary thresholding strategies.

Although our target scenario is computer-assisted translation, which calls for solutions to present human translators with useful MT suggestions (easier to correct than to rewrite from scratch), our method aims to produce reliable datasets suitable for any application where a sharp distinction between “good” and “bad” translations is required.

To promote research on QE as a binary classification task and facilitate the development of binary QE applications, the three annotated datasets have been combined in a single resource, called BinQE, freely available at: <http://hlt.fbk.eu/technologies/binqe>.

To the best of our knowledge, these represent the first freely available QE corpora empirically (and reliably) annotated with binary labels.

5. Acknowledgements

This work has been partially supported by the EC-funded project MateCat (ICT-2011.4.2-287688).

6. References

- Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: a Method for Measuring Machine Translation Confidence. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 211–219. The Association for Computer Linguistics.
- Frédéric Blain, Holger Schwenk, and Jean Senellart. 2012. Incremental Adaptation Using Translation Information and Post-Editing Analysis. In *International Workshop on Spoken Language Translation*, pages 234–241, Hong-Kong (China).
- Ondrej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Eighth Workshop on Statistical Machine Translation, WMT-2013*, pages 1–44, Sofia, Bulgaria.
- Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation (WMT’12)*, pages 10–51, Montréal, Canada.
- Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May.

- Chien-Ying Chen, Jen-Yuan Yeh, and Hao-Ren Ke. 2010. Plagiarism Detection using ROUGE and WordNet. *Journal of Computing*, 2(3).
- 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 on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.
- Philip Koehn. 2005. Europarl: a Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X*, pages 79–86, Phuket, Thailand.
- Maarit Koponen. 2012. Comparing Human Perceptions of Post-editing Effort with Post-editing Operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 181–190. Association for Computational Linguistics.
- Michael Lesk. 1986. Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th annual international conference on Systems documentation (SIGDOC86)*.
- Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the ACL workshop on Text Summarization Branches Out.*, pages 74–81, Barcelona, Spain.
- Alessia Mammone, Marco Turchi, and Nello Cristianini. 2009. Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(3):283–289.
- Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012. Match without a Referee: Evaluating MT Adequacy without Reference Translations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation, WMT '12*, pages 171–180, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.
- M. Porter. 2001. Snowball: A language for stemming algorithms.
- Marion Potet, Laurent Besacier, and Hervé Blanchon. 2010. The LIG Machine Translation System for WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 161–166, Uppsala, Sweden, July. Association for Computational Linguistics.
- Marion Potet, Emmanuelle Esperana-Rodier, Laurent Besacier, and Hervé Blanchon. 2012. Collection of a Large Database of French-English SMT Output Corrections. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- Martin Potthast, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. 2010. Overview of the 2nd International Competition on Plagiarism Detection. *Notebook Papers of CLEF*, 10.
- Matthew Snover, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. 2009. Fluency, Adequacy, or HTER?: Exploring Different Human Judgments with a Tunable MT Metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation, StatMT '09*, pages 259–268.
- Radu Soricut and Abdessamad Echihabi. 2010. TrustRank: Inducing Trust in Automatic Translations via Ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, pages 612–621, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT'09)*, pages 28–35, Barcelona, Spain.
- Lucia Specia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. pages 73–80.
- Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.
- Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the Subjectivity of Human Judgements in MT Quality Estimation. In *Proceedings of the 8th Workshop on Statistical Machine Translation (WMT'13)*, Sofia, Bulgaria.
- Marco Turchi, Antonios Anastasopoulos, José G.C. de Souza, and Matteo Negri. 2014. Adaptive Quality Estimation for Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL '14)*. Association for Computational Linguistics.