Reducing Annotation Effort for Quality Estimation via Active Learning

Daniel Beck and **Lucia Specia** and **Trevor Cohn**
Department of Computer Science
University of Sheffield
Sheffield, United Kingdom
{debeck1,l.specia,t.cohn}@sheffield.ac.uk

Abstract

Quality estimation models provide feedback on the quality of machine translated texts. They are usually trained on human-annotated datasets, which are very costly due to its task-specific nature. We investigate active learning techniques to reduce the size of these datasets and thus annotation effort. Experiments on a number of datasets show that with as little as 25% of the training instances it is possible to obtain similar or superior performance compared to that of the complete datasets. In other words, our active learning query strategies can not only reduce annotation effort but can also result in better quality predictors.

1 Introduction

The purpose of machine translation (MT) quality estimation (QE) is to provide a quality prediction for new, unseen machine translated texts, without relying on reference translations (Blatz et al., 2004; Specia et al., 2009; Callison-Burch et al., 2012). This task is usually addressed with machine learning models trained on datasets composed of source sentences, their machine translations, and a quality label assigned by humans. A common use of quality predictions is the decision between post-editing a given machine translated sentence and translating its source from scratch, based on whether its post-editing effort is estimated to be lower than the effort of translating the source sentence.

Since quality scores for the training of QE models are given by human experts, the annotation process is costly and subject to inconsistencies due to the subjectivity of the task. To avoid inconsistencies because of disagreements among annotators, it is often recommended that a QE model is trained for each translator, based on labels given by such a translator (Specia, 2011). This further increases the annotation costs because different datasets are needed for different tasks. Therefore, strategies to reduce the demand for annotated data are needed. Such strategies can also bring the possibility of selecting data that is less prone to inconsistent annotations, resulting in more robust and accurate predictions.

In this paper we investigate Active Learning (AL) techniques to reduce the size of the dataset while keeping the performance of the resulting QE models. AL provides methods to select informative data points from a large pool which, if labelled, can potentially improve the performance of a machine learning algorithm (Settles, 2010). The rationale behind these methods is to help the learning algorithm achieve satisfactory results from only on a subset of the available data, thus incurring less annotation effort.

2 Related Work

Most research work on QE for machine translation is focused on feature engineering and feature selection, with some recent work on devising more reliable and less subjective quality labels. Blatz et al. (2004) present the first comprehensive study on QE for MT: 91 features were proposed and used to train predictors based on an automatic metric (e.g. NIST (Doddington, 2002)) as the quality label. Quirk (2004) showed that small datasets manually annotated by humans for quality can result in models that outperform those trained on much larger, automatically labelled sets.

Since quality labels are subjective to the annotators' judgements, Specia and Farzindar (2010) evaluated the performance of QE models using HTER (Snover et al., 2006) as the quality score, i.e., the edit distance between the MT output and its post-edited version. Specia (2011) compared the performance of models based on labels for

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 543–548,
Sofia, Bulgaria, August 4-9 2013.

post-editing effort, post-editing time, and HTER.

In terms of learning algorithms, by and large most approaches use Support Vector Machines, particularly regression-based approaches. For an overview on various feature sets and machine learning algorithms, we refer the reader to a recent shared task on the topic (Callison-Burch et al., 2012).

Previous work use supervised learning methods ("passive learning" following the AL terminology) to train QE models. On the other hand, AL has been successfully used in a number of natural language applications such as text classification (Lewis and Gale, 1994), named entity recognition (Vlachos, 2006) and parsing (Baldridge and Osborne, 2004). See Olsson (2009) for an overview on AL for natural language processing as well as a comprehensive list of previous work.

3 Experimental Settings

3.1 Datasets

We perform experiments using four MT datasets manually annotated for quality:

English-Spanish (*en-es*): $2,254$ sentences translated by Moses (Koehn et al., 2007), as provided by the WMT12 Quality Estimation shared task (Callison-Burch et al., 2012). Effort scores range from 1 (too bad to be post-edited) to 5 (no post-editing needed). Three expert post-editors evaluated each sentence and the final score was obtained by a weighted average between the three scores. We use the default split given in the shared task: $1,832$ sentences for training and 432 for test.

French-English (*fr-en*): $2,525$ sentences translated by Moses as provided in Specia (2011), annotated by a single translator. Human labels indicate post-editing effort ranging from 1 (too bad to be post-edited) to 4 (little or no post-editing needed). We use a random split of 90% sentences for training and 10% for test.

Arabic-English (*ar-en*): $2,585$ sentences translated by two state-of-the-art SMT systems (denoted *ar-en-1* and *ar-en-2*), as provided in (Specia et al., 2011). A random split of 90% sentences for training and 10% for test is used. Human labels indicate the adequacy of the translation ranging from 1 (completely inadequate) to 4 (adequate). These datasets were annotated by two expert translators.

3.2 Query Methods

The core of an AL setting is how the learner will gather new instances to add to its training data. In our setting, we use a pool-based strategy, where the learner queries an instance pool and selects the best instance according to an informativeness measure. The learner then asks an "oracle" (in this case, the human expert) for the true label of the instance and adds it to the training data.

Query methods use different criteria to predict how informative an instance is. We experiment with two of them: Uncertainty Sampling (US) (Lewis and Gale, 1994) and Information Density (ID) (Settles and Craven, 2008). In the following, we denote $M(x)$ the query score with respect to method M.

According to the US method, the learner selects the instance that has the highest labelling variance according to its model:

$$US(x) = Var(y|x)$$

The ID method considers that more dense regions of the query space bring more useful information, leveraging the instance uncertainty and its similarity to all the other instances in the pool:

$$ID(x) = Var(y|x) \times \left(\frac{1}{U} \sum_{u=1}^{U} sim(x, x^{(u)}) \right)^{\beta}$$

The β parameter controls the relative importance of the density term. In our experiments, we set it to 1, giving equal weights to variance and density. The U term is the number of instances in the query pool. As similarity measure $sim(x, x^{(u)})$, we use the cosine distance between the feature vectors. With each method, we choose the instance that maximises its respective equation.

3.3 Experiments

To build our QE models, we extracted the 17 features used by the baseline approach in the WMT12 QE shared task.[1] These features were used with a Support Vector Regressor (SVR) with radial basis function and fixed hyperparameters (C=5, γ=0.01, ϵ=0.5), using the Scikit-learn toolkit (Pedregosa et al., 2011). For each dataset and each query method, we performed 20 active learning simulation experiments and averaged the results. We

[1] We refer the reader to (Callison-Burch et al., 2012) for a detailed description of the feature set, but this was a very strong baseline, with only five out of 19 participating systems outperforming it.

started with 50 randomly selected sentences from the training set and used all the remaining training sentences as our query pool, adding one new sentence to the training set at each iteration.

Results were evaluated by measuring Mean Absolute Error (MAE) scores on the test set. We also performed an "oracle" experiment: at each iteration, it selects the instance that minimises the MAE on the test set. The oracle results give an upper bound in performance for each test set.

Since an SVR does not supply variance values for its predictions, we employ a technique known as *query-by-bagging* (Abe and Mamitsuka, 1998). The idea is to build an ensemble of N SVRs trained on sub-samples of the training data. When selecting a new query, the ensemble is able to return N predictions for each instance, from where a variance value can be inferred. We used 20 SVRs as our ensemble and 20 as the size of each training sub-sample.[2] The variance values are then used as-is in the case of US strategy and combined with query densities in case of the ID strategy.

4 Results and Discussion

Figure 1 shows the learning curves for all query methods and all datasets. The "random" curves are our baseline since they are equivalent to passive learning (with various numbers of instances). We first evaluated our methods in terms of how many instances they needed to achieve 99% of the MAE score on the full dataset. For three datasets, the AL methods significantly outperformed the random selection baseline, while no improvement was observed on the *ar-en-1* dataset. Results are summarised in Table 1.

The learning curves in Figure 1 show an interesting behaviour for most AL methods: some of them were able to yield lower MAE scores than models trained on the full dataset. This is particularly interesting in the *fr-en* case, where both methods were able to obtain better scores using only ∼25% of the available instances, with the US method resulting in 0.03 improvement. The random selection strategy performs surprisingly well (for some datasets it is better than the AL strategies with certain number of instances), providing extra evidence that much smaller annotated

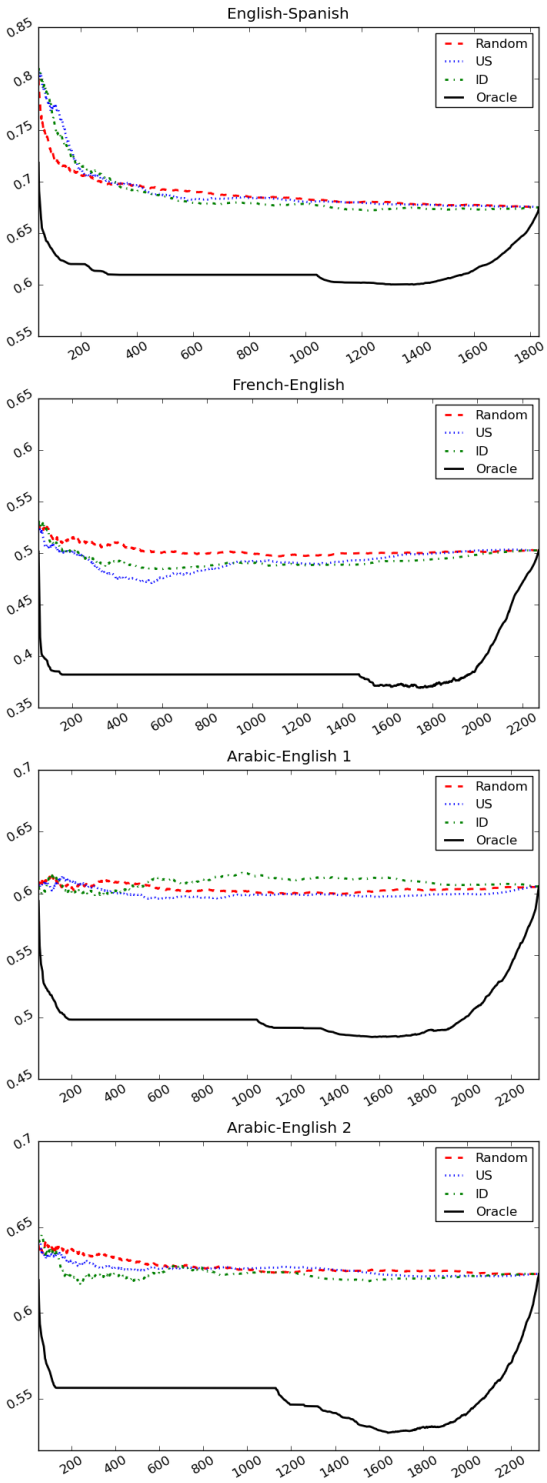

Figure 1: Learning curves for different query selection strategies in the four datasets. The horizontal axis shows the number of instances in the training set and the vertical axis shows MAE scores.

[2]We also tried sub-samples with the same size of the current training data but this had a large impact in the query methods running time while not yielding significantly better results.

	US		ID		Random		Full dataset
	#instances	MAE	#instances	MAE	#instances	MAE	
en-es	959 (52%)	0.6818	**549 (30%)**	0.6816	1079 (59%)	0.6818	0.6750
fr-en	**79 (3%)**	0.5072	134 (6%)	0.5077	325 (14%)	0.5070	0.5027
ar-en-1	51 (2%)	0.6067	51 (2%)	0.6052	51 (2%)	0.6061	0.6058
ar-en-2	209 (9%)	0.6288	**148 (6%)**	0.6289	532 (23%)	0.6288	0.6290

Table 1: Number (proportion) of instances needed to achieve 99% of the performance of the full dataset. Bold-faced values indicate the best performing datasets.

	Best MAE US			Best MAE ID			Full dataset
	#instances	MAE US	MAE Random	#instances	MAE ID	MAE Random	
en-es	1832 (100%)	0.6750	0.6750	1122 (61%)	**0.6722**	0.6807	0.6750
fr-en	559 (25%)	**0.4708**	0.5010	582 (26%)	0.4843	0.5008	0.5027
ar-en-1	610 (26%)	**0.5956**	0.6042	351 (15%)	0.5987	0.6102	0.6058
ar-en-2	1782 (77%)	0.6212	0.6242	190 (8%)	**0.6170**	0.6357	0.6227

Table 2: Best MAE scores obtained in the AL experiments. For each method, the first column shows the number (proportion) of instances used to obtain the best MAE, the second column shows the MAE score obtained and the third column shows the MAE score for random instance selection at the same number of instances. The last column shows the MAE obtained using the full dataset. Best scores are shown in bold and are significantly better (paired t-test, $p < 0.05$) than both their randomly selected counterparts and the full dataset MAE.

datasets than those used currently can be sufficient for machine translation QE.

The best MAE scores achieved for each dataset are shown in Table 2. The figures were tested for significance using pairwise t-test with 95% confidence,[3] with bold-faced values in the table indicating significantly better results.

The lower bounds in MAE given by the oracle curves show that AL methods can indeed improve the performance of QE models: an ideal query method would achieve a very large improvement in MAE using fewer than 200 instances in all datasets. The fact that different datasets present similar oracle curves suggests that this is not related for a specific dataset but actually a common behaviour in QE. Although some of this gain in MAE may be due to overfitting to the test set, the results obtained with the *fr-en* and *ar-en-2* datasets are very promising, and therefore we believe that it is possible to use AL to improve QE results in other cases, as long as more effective query techniques are designed.

5 Further analysis on the oracle behaviour

By analysing the oracle curves we can observe another interesting phenomenon which is the rapid increase in error when reaching the last ∼200 instances of the training data. A possible explanation for this behaviour is the existence of erroneous, inconsistent or contradictory labels in the datasets. Quality annotation is a subjective task by nature, and it is thus subject to noise, e.g., due to misinterpretations or disagreements. Our hypothesis is that these last sentences are the most difficult to annotate and therefore more prone to disagreements.

To investigate this phenomenon, we performed an additional experiment with the *en-es* dataset, the only dataset for which multiple annotations are available (from three judges). We measure the Kappa agreement index (Cohen, 1960) between all pairs of judges in the subset containing the first 300 instances (the 50 initial random instances plus 250 instances chosen by the oracle). We then measured Kappa in windows of 300 instances until the last instance of the training set is selected by the oracle method. We also measure variances in sentence length using windows of 300 instances. The idea of this experiment is to test whether sentences that are more difficult to annotate (because of their length or subjectivity, generating more disagreement between the judges) add noise to the dataset.

The resulting Kappa curves are shown in Figure 2: the agreement between judges is high for the initial set of sentences selected, tends to decrease until it reaches ∼1000 instances, and then starts to increase again. Figure 3 shows the results for source sentence length, which follow the same trend (in a reversed manner). Contrary to our hy-

[3] We took the average of the MAE scores obtained from the 20 runs with each query method for that.

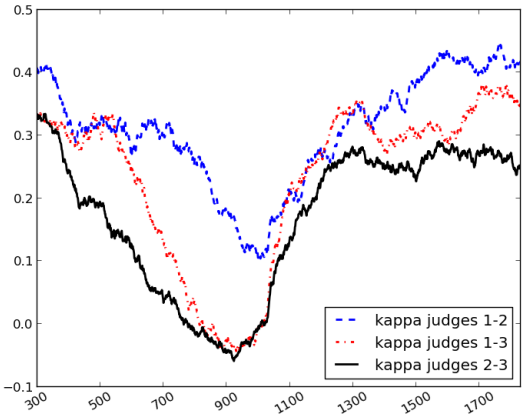

Figure 2: Kappa curves for the en-es dataset. The horizontal axis shows the number of instances and the vertical axis shows the kappa values. Each point in the curves shows the kappa index for a window containing the last 300 sentences chosen by the oracle.

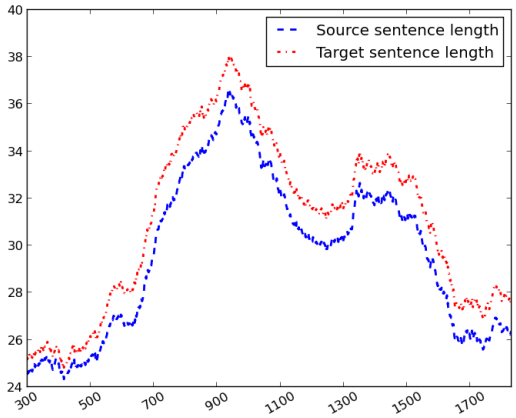

Figure 3: Average source and target sentence lengths for the *en-es* dataset. The horizontal axis shows the number of instances and the vertical axis shows the length values. Each point in the curves shows the average length for a window containing the last 300 sentences chosen by the oracle.

pothesis, these results suggest that the most difficult sentences chosen by the oracle are those in the middle range instead of the last ones. If we compare this trend against the oracle curve in Figure 1, we can see that those middle instances are the ones that do not change the performance of the oracle.

The resulting trends are interesting because they give evidence that sentences that are difficult to annotate do not contribute much to QE performance (although not hurting it either). However, they do not confirm our hypothesis about the oracle behaviour. Another possible source of disagreement is the feature set: the features may not be discriminative enough to distinguish among different instances, i.e., instances with very similar features but different labels might be genuinely different, but the current features are not sufficient to indicate that. In future work we plan to further investigate this by hypothesis by using other feature sets and analysing their behaviour.

6 Conclusions and Future Work

We have presented the first known experiments using active learning for the task of estimating machine translation quality. The results are promising: we were able to reduce the number of instances needed to train the models in three of the four datasets. In addition, in some of the datasets active learning yielded significantly better models using only a small subset of the training instances.

The oracle results give evidence that it is possible to go beyond these encouraging results by employing better selection strategies in active learning. In future work we will investigate more advanced query techniques that consider features other than variance and density of the data points. We also plan to further investigate the behaviour of the oracle curves using not only different feature sets but also different quality scores such as HTER and post-editing time. We believe that a better understanding of this behaviour can guide further developments not only for instance selection techniques but also for the design of better quality features and quality annotation schemes.

Acknowledgments

This work was supported by funding from CNPq/Brazil (No. 237999/2012-9, Daniel Beck) and from the EU FP7-ICT QTLaunchPad project (No. 296347, Lucia Specia).

References

Naoki Abe and Hiroshi Mamitsuka. 1998. Query learning strategies using boosting and bagging. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 1–9.

Jason Baldridge and Miles Osborne. 2004. Active learning and the total cost of annotation. In *Proceedings of EMNLP*, pages 9–16.

John Blatz, Erin Fitzgerald, and George Foster. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th Conference on Computational Linguistics*, pages 315–321.

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 7th Workshop on Statistical Machine Translation*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 128–132.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 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.

David D. Lewis and Willian A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–10.

Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing. Technical report.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Duborg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Chris Quirk. 2004. Training a sentence-level machine translation confidence measure. In *Proceedings of LREC*, pages 825–828.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079.

Burr Settles. 2010. Active learning literature survey. Technical report.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.

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

Lucia Specia, M Turchi, Zhuoran Wang, and J Shawe-Taylor. 2009. Improving the confidence of machine translation quality estimates. In *Proceedings of MT Summit XII*.

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

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of EAMT*.

Andreas Vlachos. 2006. Active annotation. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining at EACL*.