Confidence-based Active Learning Methods for Machine Translation

Varvara Logacheva
University of Sheffield
Sheffield, United Kingdom
v.logacheva@sheffield.ac.uk

Lucia Specia
University of Sheffield
Sheffield, United Kingdom
l.specia@sheffield.ac.uk

Abstract

The paper presents experiments with active learning methods for the acquisition of training data in the context of machine translation. We propose a confidence-based method which is superior to the state-of-the-art method both in terms of quality and complexity. Additionally, we discovered that oracle selection techniques that use real quality scores lead to poor results, making the effectiveness of confidence-driven methods of active learning for machine translation questionable.

1 Introduction

Active learning (AL) is a technique for the automatic selection of data which is most useful for model building. In the context of machine translation (MT), AL is particularly important as the acquisition of data often has a high cost, i.e. new source texts need to be translated manually. Thus it is beneficial to select for manual translation sentences which can lead to better translation quality.

The majority of AL methods for MT is based on the (dis)similarity of sentences with respect to the training data, with particular focus on domain adaptation. Eck et al. (2005) suggest a TF-IDF metric to choose sentences with words absent in the training corpus. Ambati et al. (2010) propose a metric of informativeness relying on unseen n-grams.

Bloodgood and Callison-Burch (2010) use n-gram frequency and coverage of the additional data as selection criteria. Their technique solicits translations for phrases instead of entire sentences, which saves user effort and leads to quality improvements even if the initial dataset is already sizeable.

A recent trend is to select source sentences based on an estimate of the quality of their translation by a baseline MT system. It is assumed that if a sentence has been translated well with the existing data, it will not contribute to improving the translation quality. If however a sentence has been translated erroneously, it might have words or phrases that are absent or incorrectly represented. Haffari et al. (2009) train a classifier to define the sentences to select. The classifier uses a set of features of the source sentences and their automatic translations: n-grams and phrases frequency, MT model score, etc. Ananthakrishnan et al. (2010) build a pairwise classifier that ranks sentences according to the proportion of n-grams they contain that can cause errors. For quality estimation, Banerjee et al. (2013) train language models of well and badly translated sentences. The usefulness of a sentence is measured as the difference of its perplexities in these two language models.

In this research we also explore a quality-based AL technique. Compared to its predecessors, our method is based on a more complex and therefore potentially more reliable quality estimation framework. It uses wider range of features, which go beyond those used in previous work, covering information from both source and target sentences.

Another important novel feature in our work is the addition of real post-editions to the MT training data, as opposed to simulated post-editions (human reference translations) as in previous work on AL for MT. As we show in section 3.2, adding post-editions leads to superior translation quality improvements. Additionally, this is a suitable solution for "human in the loop" settings, as post-editing automatically translated sentences tends to be faster and easier than translation from scratch (Koehn and Haddow, 2009). Also, different from previous work, we do not focus on domain adaptation: our experiments involve only in-domain data.

Compared to previous work on confidence-driven AL, our approach has led to better results, but these proved to be highly dependent on a sentence length bias. However, an oracle-based selec-

Workshop on Humans and Computer-assisted Translation, pages 78–83,
Gothenburg, Sweden, 26 April 2014.

tion using true quality scores has not been shown to perform well. This indicates that the usefulness of quality scores as AL selection criterion in the context of MT needs to be further investigated.

2 Active selection strategy

Our AL sentence selection strategy relies on quality estimation (QE). QE is aimed at predicting the quality of a translated text (in this case, a sentence) without resorting to reference translations. It considers features of the source and machine translated texts, and an often small number (a few hundreds) of examples of translations labelled for quality by humans to train a machine learning algorithm to predict such quality labels for new data.

We use the open source QE framework QuEst (Specia et al., 2013). In our settings it was trained to predict an HTER score (Snover et al., 2006) for each sentence, i.e., the edit distance between the automatic translation and its human post-edited version. QuEst can extract a wide range of features. In our experiments we use only the 17 so-called *baseline features*, which have been shown to perform well in evaluation campaigns (Bojar et al., 2013): number of tokens in sentences, average token length, language model probabilities for source and target sentences, average number of translations per source word, percentage of higher and lower frequency n-grams in source sentence based on MT training corpus, number of punctuation marks in source and target sentences.

Similarly to Ananthakrishnan et al. (2010), we assume that the most useful sentences are those that lead to larger translation errors. However, instead of looking at the n-grams that caused errors — a very sparse indicator requiring significantly larger amounts of training data, we account for errors in a more general way: the (QuEst predicted) percentage of edits (HTER) that would be necessary to transform the MT output into a correct sentence.

3 Experiments and results

3.1 Datasets and MT settings

For the AL data selection experiment, two datasets are necessary: parallel sentences to train an initial, baseline MT system, and an additional pool of parallel sentences to select from. Our goal was to study potential improvements in the baseline MT system in a realistic "human in the loop" scenario, where source sentences are translated by the baseline system and post-edited by humans before they are added to the system. As it has been shown in (Potet et al., 2012), post-editions tend to be closer to source sentences than freely created translations. One of our research questions was to investigate whether they would be more useful to improve MT quality.

We chose the biggest corpus with machine translations and post-editions available to date: the LIG French–English post-editions corpus (Potet et al., 2012). It contains 10,881 quadruples of the type: *<source sentence, reference translation, automatic translation, post-edited automatic translation>*. Out of these, we selected 9,000 as the pool to be added to be baseline MT system, and the remaining 1,881 to train the QE system for the experiments with AL. For QE training, we use the HTER scores between MT and its post-edited version as computed by the TERp tool.[1]

We use the Moses toolkit with standard settings[2] to build the (baseline) statistical MT systems. As training data, we use the French–English News Commentary corpus released by the WMT13 shared task (Bojar et al., 2013). For the AL experiments, the size of the pool of additional data (10,000) poses a limitation. To examine improvements obtained by adding fractions of up to only 9,000 sentences, we took a small random subset of the WMT13 data for these experiments (Table 1). Although these figures may seem small, the settings are realistic for many language pairs and text domains where larger data sets are simply not available.

We should also note that all the data used in our experiments belongs to the same domain: the LIG SMT system which produced sentences for the post-editions corpus was trained on Europarl and News commentary datasets (Potet et al., 2010), but the post-edited sentences themselves were taken from *news* test sets released for WMT shared tasks in different years. Our baseline system is trained on a fraction of the *news* commentary corpus. Finally, we tune and test all our systems on WMT shared task *news* news datasets (those which do not overlap with the post-editions corpus).

[1] http://www.umiacs.umd.edu/~snover/terp/

[2] http://www.statmt.org/moses/?n=Moses.Baseline

Corpora	Size (sentences)
Initial data (baseline MT system)	
Training - subset of News Commentary corpus	10,000
Tuning - WMT newstest-2012	3,000
Test - WMT newstest-2013	3,000
Additional data (AL data)	
Post-editions corpus:	10,881
- **Training QE** system	1,881
- **AL pool**	9,000

Table 1: Datasets

3.2 Post-editions versus references

In order to compare the impact of post-editions and reference translations on MT quality, we added these two variants of translations to baseline MT systems of different sizes, including the entire News Commentary corpus. The figures for BLEU (Papineni et al., 2002) scores in Table 2 show that adding post-editions results in significantly better quality than adding the same number of reference translations[3]. This effect can be seen even when the additional data corresponds to only a small fraction of the training data.

In addition, it does not seem to matter which MT system produced the translations which were then post-edited in the post-edition corpus. Even if the output of a third-party system was used (as in our case), it improves the quality of machine translations for unseen data. We assume that since post-editions tend to be closer to original sentences than free translations (Potet et al., 2012), they generally help produce better source-target alignments, leading to the extraction of good quality phrases.

Baseline corpus (sentences)	Results (BLEU)		
	Baseline	**Ref**	**PE**
150,000	22.41	22.95	23.21
50,000	20.22	20.91	22.01
10,000	15.09	18.65	20.44

Table 2: Influence of post-edited and reference translations on MT quality. **Ref**: baseline system with added free references, **PE**: baseline system with added post-editions.

[3]These systems use the whole post-editions set (10,881 sentences) as opposed to 9,000-sentence subset which we use further in our AL experiments. Therefore the figures reported in this table are higher than those in subsequent sections.

3.3 AL settings

The experimental settings for all methods are as follows. First, a baseline MT system is trained. Then a batch of 1,000 sentences is selected from the data pool with an AL strategy, and the selected data is removed from the pool. The MT system is rebuilt using a concatenation of the initial training data and the new batch. The process is repeated until the pool is empty, with subsequent steps using the MT system trained on the previous step as a baseline. The performance of each MT system is measured in terms of BLEU scores. We use the following AL strategies:

- **QuEst**: our method described in section 2.
- **Random**: random selection of sentences.
- **HTER**: oracle-based selection based on true HTER scores of sentences in the pool, instead of the QuEst estimated HTER scores.
- **Ranking**: AL strategy described in (Ananthakrishnan et al., 2010) for comparison.

3.4 AL results

Our initial results in Figure 1 show that our selection strategy (**QuEst**) consistently outperforms the **Random** selection baseline.

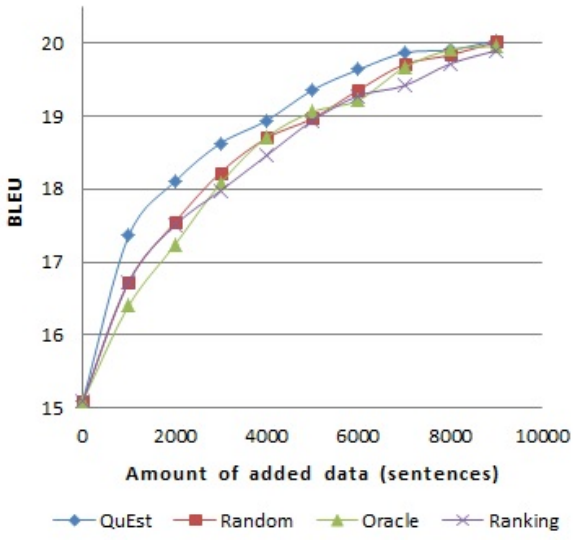

Figure 1: Performance of MT systems enhanced with data selected by different AL strategies

In comparison with previous work, we found that the error-based **Ranking** strategy performs closely to **Random** selection, although (Ananthakrishnan et al., 2010) reports it to be better.

Compared to **QuEst**, we believe the lower figures of the **Ranking** strategy are due to the fact that the latter considers features of only one type (source n-grams), whereas QuEst uses a range of different features of the source and translation sentences.

Interestingly, the **Oracle** method underperforms our QE-based method, although we expected the use of real HTER scores to be more effective. In order to understand the reasons behind such behaviour, we examined the batches selected by **QuEst** and **Oracle** strategies more closely. We found that the distribution of sentence lengths in batches by the two strategies is very different (see Figure 2). While in batches selected by **QuEst** the average sentence length steadily decreases as more data is added, in **Oracle** batches the average length was almost uniform for all batches, except the first one, which contains shorter sentences.

This is explained by HTER formulation: HTER is computed as the number of edits over the sentence length, and therefore in shorter sentences every edit is given more weight. For example, the HTER score of a 5-word sentence with one error is 0.2, whereas a sentence of 20 words with the same single error has a score of 0.05. However, it is doubtful that the former sentence will be more useful for an MT system than the latter. Regarding the nature of length bias in the predictions done by **QuEst** system, sentence length is used there as a feature, and longer sentences tend to be estimated as having higher HTER scores (i.e., lower translation quality).

Therefore, sentences with the highest HTER may not actually be the most useful, which makes the **Oracle** strategy inferior to **QuEst**. Moreover, longer sentences chosen by our strategy simply provide more data, so their addition might be more useful even regardless of the amount of errors.

This seems to indicate that the success of our strategy might not be related to the quality of the translations only, but to their length. Another possibility is that sentences selected by **QuEst** might have more errors, which means that they can contribute more to the MT system.

3.5 Additional experiments

In order to check the two hypotheses put forward in the previous section, we conduct two other sets of AL experiments: (i) a selection strategy that chooses longer sentences first (denoted as **Length**)

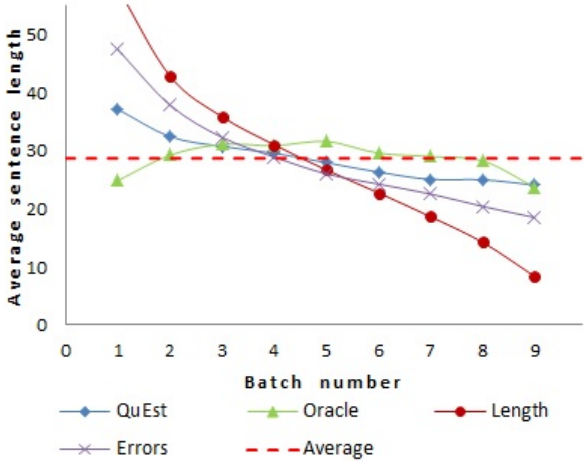

Figure 2: Number of words in batches selected by different AL strategies

and (ii) a selection strategy that chooses sentences with larger numbers of errors first (**Errors**).

Figure 3 shows that a simple length-based strategy yields better results than any of the other tested strategies. Therefore, in cases when the corpus has sufficient variation in sentence length, length-based selection might perform at least as well as other more sophisticated criteria. The experiments with confidence-based selection described in (Ananthakrishnan et al., 2010) were free of this length bias, as sentences much longer or shorter than average were deliberately filtered out.

Interestingly, results for the **Errors** strategy are slightly worse than those for **QuEst**, although the former is guaranteed to choose sentences with the largest number of errors and has even stronger length bias than **QuEst** (see figure 2). Therefore, the reasons hypothesised to be behind the superiority of **QuEst** over **Oracle** (longer sentences and larger number of errors) are actually not the only factors that influence the quality of an AL strategy.

3.6 Length-independent results

Despite the success of the length-based strategy, we do not believe that it is enough for an effective AL technique. First of all, the experiment with the **Errors** strategy demonstrated that more data does not always lead to better results. Furthermore, our aim is to reduce the translator's effort in cases when the additional data needs to be translated or post-edited manually. However, longer sentences usually take more time to translate or edit, so choosing the longest sentences from a pool of sentences will not reduce translator's effort.

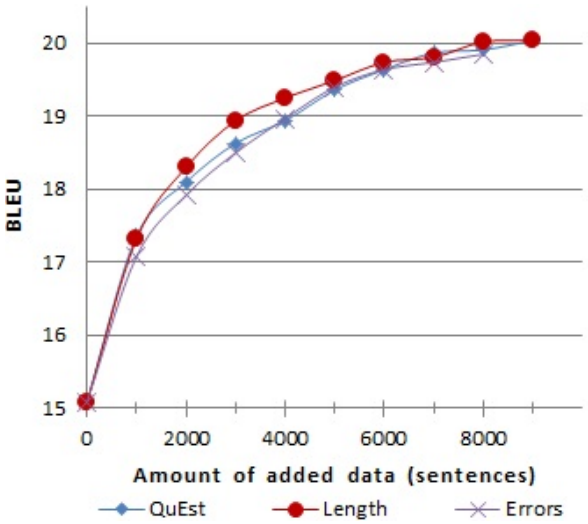

Figure 3: Comparison of our QuEst-based selection with a length-based selection

Therefore, we would like to study the effectiveness of our strategy by isolating the sentence length bias. One option is to filter out long sentences, as it was done in (Ananthakrishnan et al., 2010). However, our pool is already too small. Therefore, we plot the performance improvements with respect to training data size in words, instead of sentences. As it was already noted by Bloodgood and Callison-Burch (2010), measuring the amount of added data in sentences can significantly contort the real annotation cost (the cost of acquisition of new translations). So we switch to length-independent representation.

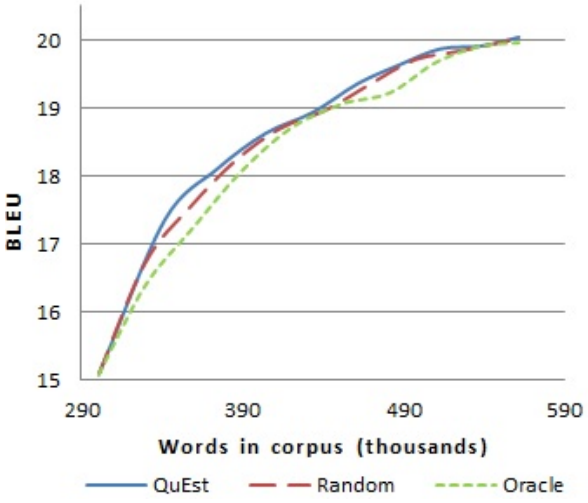

Figure 4: Active learning quality plotted with respect to data size in words: **QuEst** vs **Oracle** strategies.

Figure 4 shows that the **Oracle** strategy in

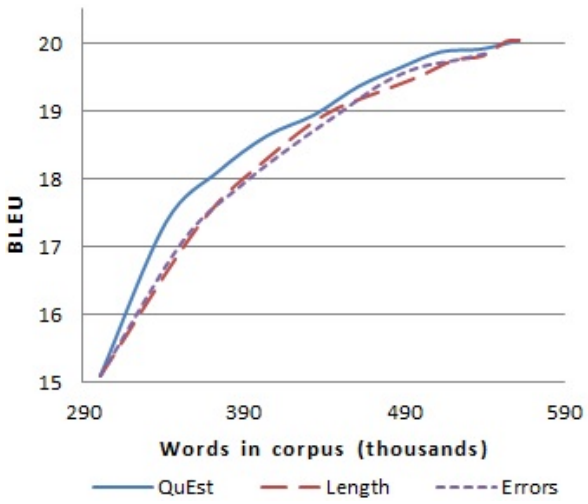

Figure 5: AL quality plotted with respect to data size in words: **QuEst** vs **Length** and **Errors** strategies.

length-independent representation can still be seen to perform worse than both our strategy and random selection. Results of **Length** and **Error** strategies (plotted separately in figure 5 for readability) are very close and both underperform our **QuEst**-based strategy and random selection of data.

Here our experience echoes the results of (Mohit and Hwa, 2007), where the authors propose the idea of *difficult to translate phrases*. It is assumed that extending an MT system with phrases that can cause difficulties during translation is more effective than simply adding new data and re-building the system. Due to the lack of time and human annotators, the authors extracted difficult phrases automatically using a set of features: alignment features, syntactic features, model score, etc. Conversely, we had the human-generated information on what segments have been translated incorrectly. We assumed that the use of this knowledge as part of our AL strategy would give us an upper bound for our AL method results. However, it turned out that prediction based on multiple features is more reliable than precise information on quality, which accounts for only one aspect of data.

4 Conclusions

We presented experiments with an active learning strategy for machine translation based on quality predictions. This strategy performs well compared to another quality-driven strategy and a random baseline. However, we found that it was success-

ful mostly due to its tendency to rate long sentences as having lower quality. Consequently, the AL application that chooses the longest sentences is not less successful when selecting from corpora with large variation in sentence length. A length-independent representation of the results showed that an oracle selection is less effective than our quality-based strategy, which we believe to be due to the nature of corrections and small size of the post-edition corpus. In addition to that, another oracle selection based on the amount of errors and length-based selection show poor results when displayed in length-independent mode.

We believe that the quality estimation strategy benefits from other features that reflect the usefulness of a sentence better than its HTER score and the amount of user corrections. In future work we will examine the influence of individual features of the quality estimation model (such as language model scores) as active learning selection strategy.

References

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2010. Active Learning and Crowd-Sourcing for Machine Translation. *LREC 2010: Proceedings of the seventh international conference on Language Resources and Evaluation, 17-23 May 2010, Valletta, Malta*, pages 2169–2174.

Sankaranarayanan Ananthakrishnan, Rohit Prasad, David Stallard, and Prem Natarajan. 2010. Discriminative Sample Selection for Statistical Machine Translation. *EMNLP-2010: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, October 9-11, 2010, MIT, Massachusetts, USA*, (October):626–635.

Pratyush Banerjee, Raphael Rubino, Johann Roturier, and Josef van Genabith. 2013. Quality Estimation-guided Data Selection for Domain Adaptation of SMT. *MT Summit XIV: proceedings of the fourteenth Machine Translation Summit, September 2-6, 2013, Nice, France*, pages 101–108.

Michael Bloodgood and Chris Callison-Burch. 2010. Bucking the Trend: Large-Scale Cost-Focused Active Learning for Statistical Machine Translation. *ACL 2010: the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, July 11-16, 2010*, pages 854–864.

Ondřej 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 *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Low Cost Portability for Statistical Machine Translation based on N-gram Frequency and TF-IDF. *IWSLT 2005: Proceedings of the International Workshop on Spoken Language Translation. October 24-25, 2005, Pittsburgh, PA*.

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09*.

Philipp Koehn and Barry Haddow. 2009. Interactive Assistance to Human Translators using Statistical Machine Translation Methods. *MT Summit XII: proceedings of the twelfth Machine Translation Summit, August 26-30, 2009, Ottawa, Ontario, Canada*, pages 73–80.

Behrang Mohit and Rebecca Hwa. 2007. Localization of Difficult-to-Translate Phrases. *ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation, June 23, 2007, Prague, Czech Republic*, pages 248–255.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *ACL 2002: 40th Annual Meeting of the Association for Computational Linguistics, July 2002, Philadelphia*, pages 311–318.

Marion Potet, Laurent Besacier, and Hervé Blanchon. 2010. The LIG machine translation system for WMT 2010. *ACL 2010: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 161–166.

Marion Potet, Emmanuelle Esperança-Rodier, Laurent Besacier, and Hervé Blanchon. 2012. Collection of a Large Database of French-English SMT Output Corrections. *LREC 2012: Eighth international conference on Language Resources and Evaluation, 21-27 May 2012, Istanbul, Turkey*, pages 4043–4048.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Visions for the Future of Machine Translation, August 8-12, 2006, Cambridge, Massachusetts, USA*, pages 223–231.

Lucia Specia, Kashif Shah, Jose G C de Souza, and Trevor Cohn. 2013. QuEst - A translation quality estimation framework. *ACL 2013: Annual Meeting of the Association for Computational Linguistics, Demo session, August 2013, Sofia, Bulgaria*.