A Quality-based Active Sample Selection Strategy for Statistical Machine Translation

Varvara Logacheva, Lucia Specia
Department of Computer Science, University of Sheffield
Sheffield, United Kingdom
{v.logacheva,l.specia}@sheffield.ac.uk

Abstract

This paper presents a new active learning technique for machine translation based on quality estimation of automatically translated sentences. It uses an error-driven strategy, i.e., it assumes that the more errors an automatically translated sentence contains, the more informative it is for the translation system. Our approach is based on a quality estimation technique which involves a wider range of features of the source text, automatic translation, and machine translation system compared to previous work. In addition, we enhance the machine translation system training data with post-edited machine translations of the sentences selected, instead of simulating this using previously created reference translations. We found that re-training systems with additional post-edited data yields higher quality translations regardless of the selection strategy used. We relate this to the fact that post-editions tend to be closer to source sentences as compared to references, making the rule extraction process more reliable.

Keywords: machine translation, active learning, human feedback

1. Introduction

One of the most efficient ways to improve the quality of a machine translation (MT) system is to enhance it with new training data. In many scenarios, monolingual data on either source or target languages (or both) tends to be abundant. However, parallel data has to be created by having humans translating monolingual content, which is an expensive process. Clever selection techniques to choose a subset with only the most useful sentences to translate from monolingual data can result in systems with higher quality using less training data. These techniques are usually referred to as active learning (AL) (Settles, 2010). Active learning has been extensively used in various Natural Language Processing (NLP) tasks, such as POS tagging (Ringger et al., 2007), parsing (Reichart and Rappoport, 2007), and sentiment analysis (Xiao and Guo, 2013).

Previous work has used active learning for MT. However, the majority of Active learning methods used in MT used mainly properties from source sentences. One of the most common criteria used in active sample selection for MT is the dissimilarity of a candidate sample to the existing data. Along those lines, Eck et al. (2005) suggest a TF-IDF-based metric, while Ambati et al. (2010) propose a metric of informativeness relying on unseen n-grams.

Bloodgood and Callison-Burch (2010) present a simple and yet effective method. The only criterion for sentence selection is the frequency of their n-grams in the training data. Their technique requests translations for phrases instead of complete sentences, which saves user effort and leads to marked improvement of quality even if the size of the initial training dataset is already substantial.

A recent trend is to use the estimated translation quality of a sentence as a criterion for active selection. Sentences which are likely to be translated well by the existing MT system will not be useful for its training. Conversely, sentences that are translated badly could contain words or phrases which are absent in the current translation model. One of the first methods of this type is presented in (Haffari et al., 2009). Haffari et al. (2009) define the most useful sentences with a classifier which uses a number of features such as the number of unseen phrases/n-grams, the similarity to the existing training data, and the confidence of translations. Ananthakrishnan et al. (2010) propose an error-driven method to define the most useful sentences. The idea is that the most useful sentences are those that lead to the largest number of translation errors. They learn a classifier that induces n-grams which are translated incorrectly. The classifier is then used to pick the source sentences that are likely to cause errors, i.e., likely to contain the largest number of erroneous n-grams.

We propose a new sentence selection strategy. Similarly to (Ananthakrishnan et al., 2010), the core idea of our method is to select sentences that are likely to be translated incorrectly by the MT system. However, we estimate the correctness of automatic translation of a sentence using a richer quality estimation metric, which benefits from a wider range of features. Another concern in our work was to analyse the potential of using human post-editions of machine-translated sentences as training data for the MT system. The use of post-editions in our research is two-fold. Firstly, the post-editions are used to train a quality estimation system which then generates the quality scores for new sentences. The translations of the new sentences selected for having low predicted quality are then post-edited, and those post-editions are added to the MT training data. We compare the improvements obtained by a system enhanced by post-editions with the improvements obtained by a system with additional reference translations.

According to the results of our experiments, the post-editions have much better impact on system's quality: the system with added post-editions outperforms the one with references by 1.5 BLEU points (Papineni et al., 2002). In addition, post-editing usually requires less time than translation from scratch (Koehn and Haddow, 2009), so replacing references with post-editions could improve both accuracy and speed of data acquisition.

The remainder of the paper is organised as follows. Section 2. contains the description of our selection method. In Section 3. we report the results of our experiments. Section 4. concludes the work and gives the directions of further research.

2. Selection strategy

Our active selection method is based on quality estimation (QE) of the machine translation output. Unlike standard MT quality evaluation measures, quality estimation does not require reference translations. It relies on properties of source and machine translated sentences which are used as features with a machine learning algorithm and quality labels at training time to build quality prediction models.

We use the `QuEst` toolkit (Specia et al., 2013) to build models to predicts HTER scores (Snover et al., 2006) for each sentence, i.e., estimates on the percentage of words in the sentence that would need to be corrected. The system therefore uses post-edited sentences at training time, and afterwards it predicts the HTER value for any unseen automatic translation using the properties of the source and translation sentences.

`QuEst` can extract two sets of features: (i) *black box* features, which are extracted from the source sentence and its automatic translation, such as word and n-gram statistics, POS statistics, and syntactic features; and (ii) *glass box* features, which also use internal information from the translation system. For our selection strategy we used only 17 so called *baseline features*. These are a subset of the black-box features that are known to perform well across datasets and language pairs. They are: number of tokens in source and target sentences, average token length in source sentences, LM probability for source and target sentences, average number of translations per source word, percentage of higher frequency and lower frequency n-grams in parallel corpora, number of punctuation marks in the source and target sentences.

We scored all sentences in a pool of additional data that could be added to the SMT training corpus with their HTER predictions and then ranked them according to these values so that the worse sentences (higher predicted HTER) appear at the top of the list.

3. Experiments

3.1. Settings

In our experiments we assume a common real-world scenario that is as follows. Only a small parallel dataset is available, and it is used to train a baseline MT system. A much larger pool of source language-only sentences is also available, from which we can choose batches of sentences to be translated using the method outlined in Section 2.. We then acquire a human translation (or a human post-edition of the automatic translation produced by the current MT system) for the chosen sentences and retrain the MT system using the original (small) corpus enhanced with the newly acquired parallel data.

As we have mentioned before, one of the aims of our work is to show the advantage of the use of post-editions of machine translations over reference translations, i.e. translations produced from scratch. Therefore, we conducted two sets of experiments. In the **first** set, translations of source sentences are done manually by a human translator. In the **second** set, we first translate the chosen sentences by the current MT system, and then pass them on to a human expert for post-editing. The post-editions form the target side of this new parallel dataset.

We compared our selection strategy with random selection for both post-editions and reference translations. As it has been shown in previous work, it is quite difficult to beat this random selection strategy as it simulates the natural distribution of data (Daumé III, 2007).

3.2. Data

Since datasets with references and post-edited translations already exist, we simulate the translation and post-edition of sentences. In particular, as pool for active learning we use the corpus of post-editions for French-English by (Potet et al., 2012). The corpus provides both reference translations and post-editions of machine translations for 10,881 source sentences. We use the first 1,881 sentences from this post-edition corpus to train the quality estimation system, so the pool from which we actually choose sentences to add to the MT training data consists of 9,000 sentences.

For the training of MT systems, we used the News Commentary French-English corpus, released for the WMT13 shared translation task.[1] We trained the initial (baseline) MT system on an a small subset of this corpus with 10,000 sentences, for both translation and language model building. As the development and test sets, we used the news test sets provided by WMT for the shared tasks in the years 2012 and 2013, respectively. The numbers on both corpora are outlined in Table 1.

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: Details on corpora used in our experiments.

[1] http://www.statmt.org/wmt13/

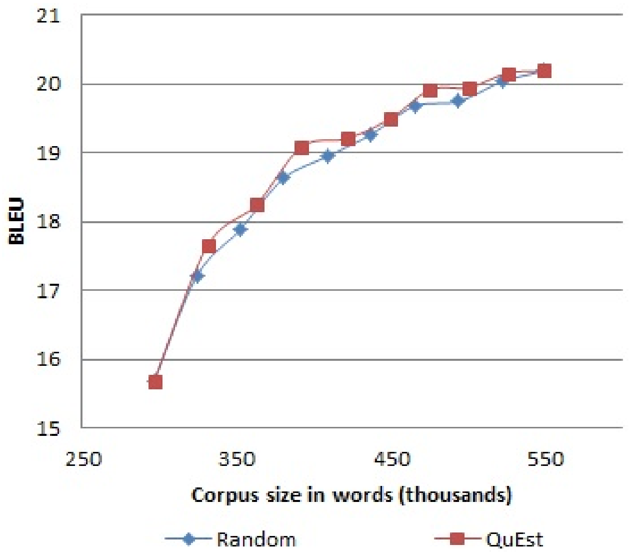

Figure 1: Performance (BLEU) improvements on the test set for MT systems re-trained based on additional data selected via active selection of **post-edited** translations

.

3.3. Results

We conducted a set of experiments to show the improvement rate of our selection strategy compared to random data selection.

At every iteration, based on quality predictions for translations of the active learning pool produced by the baseline MT system, batches of 1,000 sentences from the pool with the lowest predicted score were selected. These were added to the training data of the MT system, which was then re-trained. The selected sentences were removed from the active learning pool. The new MT system was applied to the test set, with performance (BLEU) measured. The process was repeated until the pool was empty.

Figure 1 plots the results of the experiments with added reference translations, while Figure 2, with added post-editions. Our selection strategy implies choosing complete sentences. However, the scores predicted by `QuEst` have a strong bias towards sentence length, i.e., longer sentences tend to be rated as requiring higher post-edition effort. Therefore, we show improvements in BLEU scores of the MT systems with respect to the corpus length in words, although the batches were chosen disregarding sentences length. All figures are reported based on the test set. As we can see, our error-based selection strategy results in consistent improvements in performance, and outperforms random selection when both post-editions and reference translations are added to the MT data. The improvements obtained by adding post-editions as opposed to reference translations are however substantially higher.

To highlight the difference between systems trained on added post-editions and those trained on reference translations, Figure 3 shows that adding post-editions results in higher BLEU scores than adding references for any number of added sentences. The improvement obtained for the entire set of 9,000 sentences was of **1.5 BLEU** points. The results for random selection (as opposed to **QuEst**-based selection) follow the same trend.

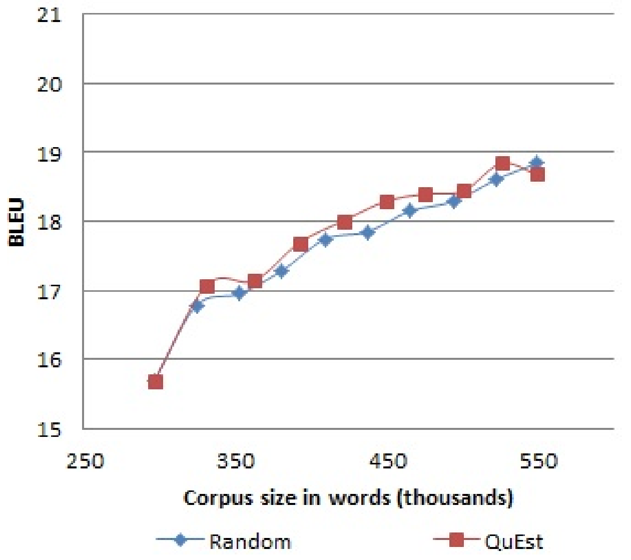

Figure 2: Performance (BLEU) improvements on the test set for MT systems re-trained based on additional data selected via active selection of **reference** translations.

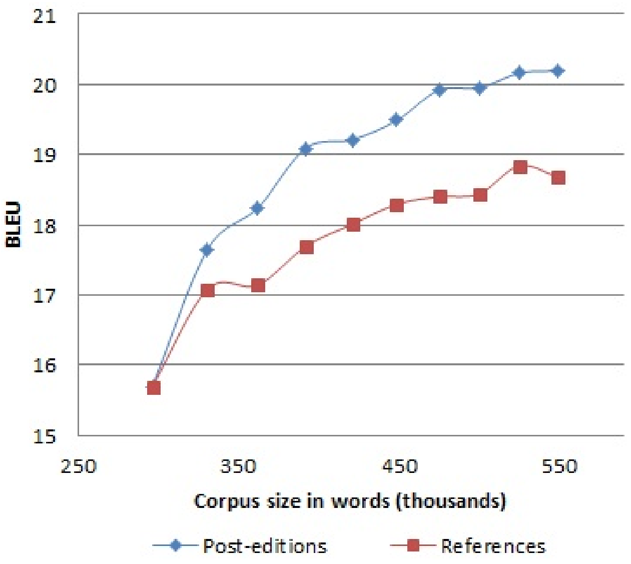

Figure 3: Comparison of the impact of post-editions and reference translations on improvements in translation quality.

3.4. Analysis

To sum up, our experiments demonstrate two main phenomena:

- MT systems trained on data selected by our error-driven active learning method yield larger improvements in translation quality than those trained on randomly selected sentences; and

- The use of automatically translated post-edited sentences as training data results in larger improvements

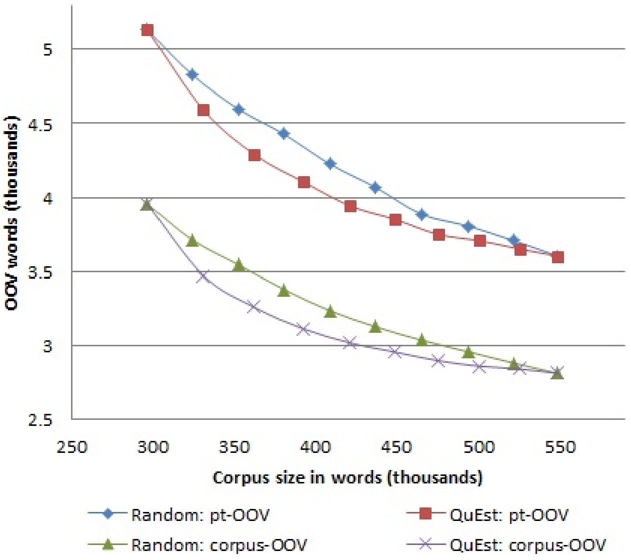

Figure 4: Comparison of reduction of OOV words for different active selection techniques.

in translation quality than the use of independently created reference translations.

In order to further understand these phenomena, we examine the number of out-of-vocabulary (OOV) words in the test set when it is translated by the different systems. Namely, we considered two parameters:

- Number of words in the test set which do not occur in training data (referred to as *corpus-OOV*);
- Number of words in the test set which do not occur in phrase table (*pt-OOV*).

These two parameters often differ as unaligned words might be omitted when extracting phrases.

Figure 4 shows how the number of OOV words in the test set naturally decreases as more data is added to MT system. This reduction is faster for both *corpus-OOV* and *pt-OOV* when our error-driven strategy is used as opposed to random selection. Therefore, QuEst seems to predict higher HTER scores for sentences that contain more uncovered words and thus we implicitly attempt to increase vocabulary variety, which is widely used in other active learning work, e.g. (Eck et al., 2005).

The comparison of OOV word rate for systems trained on post-edited translations and those trained on reference translations has also led to interesting results. Figure 5 shows that while *corpus-OOV* is almost identical for corpora containing references and post-editions, the *pt-OOV* rate for systems built from post-editions is consistently lower than that of systems built from references.

Similar *corpus-OOV* rates are expected, since the source sides of both corpora are the same (slight differences may appear because different sentences can be filtered out during the corpus cleaning step before systems are re-trained).

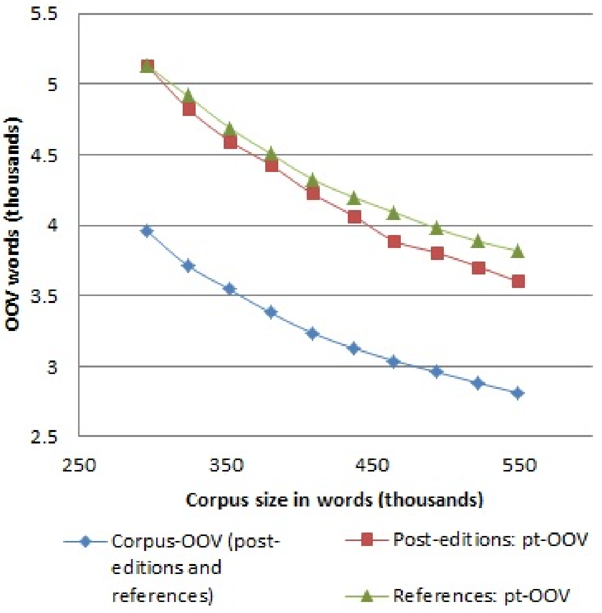

Figure 5: Comparison of the impact of post-editions and references on the OOV rate.

On the other hand, faster reduction of *pt-OOV* rate means that more phrases are extracted from the newly acquired corpus.

Overall, we can assume that post-editions contain more literal word-to-word translations and less reformulations, than reference translations. This argument is supported by research in (Potet et al., 2011). The authors report that post-editions are often closer to source sentences than their reference translations. This result means that a parallel corpus can be better for MT purposes if the target sentences are as close as possible to the literal translation of their corresponding source sentences. While this conclusion looks pretty evident as such, our experiments suggest that a simple and natural way of generating such literal translations is through the post-editing of translations produced by the underlying MT system.

We note that the post-editions used in our experiments were produced by editing an output of a third-party MT system (the LIG system (Potet et al., 2010)). Although it is built on the same type of data as our baseline system, we have only used a small fraction of the corpora, and therefore the two systems are substantially different. Nevertheless, the use of these post-editions improved the quality of our MT systems. Hence we can suggest that the actual system used to generate the translations to be post-edited is less relevant.

4. Discussion and future work

We have introduced a data selection strategy for machine translation training data collection. It selects sentences which can potentially improve the performance of an MT system, in other words, sentences which are the most useful for MT system training. We assume that the most useful sentences are those which cause more errors in a baseline system, and to judge that we look at both source and machine translated sentences.

The selection strategy is based on a quality estimation metric which predicts HTER scores. We choose sentences with the highest predicted HTER (the largest proportion of editing needed) and enhance the training data with their translations or post-editions. This strategy has been shown to outperform random selection.

We also show that an MT system with added post-edited sentences consistently outperforms and MT system with added reference translations for the same sentences. This finding suggests that we could reduce the translator's effort in creating data for active learning while getting even better improvements in the quality of the resulting MT system.

Our future research will include the comparison of our technique with other related methods, for example, the error-based techniques represented in (Ananthakrishnan et al., 2010) and (Banerjee et al., 2013). Another direction is a more in depth evaluation of our method. It will include the training of the quality estimation model on post-edited output of our baseline system, as opposed to a third party system. After retraining our MT system with a new batch of post-edited data, the quality estimation system could be retrained as well to adapt to the current state of the MT system. As a by-product of this experiment, we will be able to compare the impact of post-edited output of a particular system versus post-editions done for some other systems.

5. Acknowledgements

This work was supported by the EXPERT (EU Marie Curie ITN No. 317471) project.

6. References

Ambati, V., Vogel, S., and Carbonell, J. (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.

Ananthakrishnan, S., Prasad, R., Stallard, D., and Natarajan, P. (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*, pages 626–635.

Banerjee, P., Rubino, R., Roturier, J., and van Genabith, J. (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.

Bloodgood, M. and Callison-Burch, C. (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.

Daumé III, H. (2007). Frustratingly Easy Domain Adaptation. *ACL 2007: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, June 2007*, pages 256–263.

Eck, M., Vogel, S., and Waibel, A. (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.*

Haffari, G., Roy, M., and Sarkar, A. (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.*

Koehn, P. and Haddow, B. (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.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL02*, pages 311–318, Philadelphia.

Potet, M., Besacier, L., and Blanchon, H. (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.

Potet, M., Esperança-Rodier, E., Blanchon, H., and Besacier, L. (2011). Preliminary Experiments on Using Users' Post-Editions to Enhance a SMT System. *EAMT 2011: proceedings of the 15th conference of the European Association for Machine Translation, 30-31 May 2011, Leuven, Belgium*, pages 161–168.

Potet, M., Esperança-Rodier, E., Besacier, L., and Blanchon, H. (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.

Reichart, R. and Rappoport, A. (2007). An Ensemble Method for Selection of High Quality Parses. *ACL 2007: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, June 2007*, pages 408–415.

Ringger, E., Mcclanahan, P., Haertel, R., Busby, G., Carmen, M., Carroll, J., Seppi, K., and Lonsdale, D. (2007). Active Learning for Part-of-Speech Tagging: Accelerating Corpus Annotation. *LAW '07 Proceedings of the Linguistic Annotation Workshop, June, 2007, Prague*, pages 101–108.

Settles, B. (2010). Active Learning Literature Survey. *Computer Sciences Technical Report 1648, University of Wisconsin-Madison.*

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (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.

Specia, L., Shah, K., de Souza, J. G. C., and Cohn, T. (2013). QuEst - A translation quality estimation framework. *ACL 2013: Annual Meeting of the Association for*

Computational Linguistics, Demo session, August 2013, Sofia, Bulgaria.

Xiao, M. and Guo, Y. (2013). Online Active Learning for Cost Sensitive Domain Adaptation. *CoNLL 2013: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, August 8-9, 2013, Sofia, Bulgaria*.