Active Learning for Post-Editing Based Incrementally Retrained MT

Aswarth Dara Josef van Genabith Qun Liu John Judge Antonio Toral
School of Computing
Dublin City University
Dublin, Ireland
{adara,josef,qliu,jjudge,atoral}@computing.dcu.ie

Abstract

Machine translation, in particular statistical machine translation (SMT), is making big inroads into the localisation and translation industry. In typical workflows (S)MT output is checked and (where required) manually post-edited by human translators. Recently, a significant amount of research has concentrated on capturing human post-editing outputs as early as possible to incrementally update/modify SMT models to avoid repeat mistakes. Typically in these approaches, MT and post-edits happen sequentially and chronologically, following the way unseen data (the translation job) is presented. In this paper, we add to the existing literature addressing the question whether and if so, to what extent, this process can be improved upon by Active Learning, where input is not presented chronologically but dynamically selected according to criteria that maximise performance with respect to (whatever is) the remaining data. We explore novel (source side-only) selection criteria and show performance increases of 0.67-2.65 points TER absolute on average on typical industry data sets compared to sequential PE-based incrementally retrained SMT.

1 Introduction and Related Research

Machine Translation (MT) has evolved dramatically over the last two decades, especially since the appearance of statistical approaches (Brown et al., 1993). In fact, MT is nowadays succesfully used in the localisation and translation industry, as for many relevant domains such as technical documentation, post-editing (PE) of MT output by human translators (compared to human translation from scratch) results in notable productivity gains, as a number of industry studies have shown convincingly, e.g. (Plitt and Masselot, 2010). Furthermore, incremental retraining and update techniques (Bertoldi et al., 2013; Levenberg et al., 2010; Mathur et al., 2013; Simard and Foster, 2013) allow these PEs to be fed back into the MT model, resulting in an MT system that is continuously updated to perform better on forthcoming sentences, which should lead to a further increase in productivity.

Typically, post-editors are presented with MT output units (sentences) in the order in which input sentences appear one after the other in the translation job. Because of this, incremental MT retraining and update models based on PE outputs also proceed in the same chronological order determined by the input data. This may be suboptimal. In this paper we study the application of Active Learning (AL) to the scenario of PE MT and subsequent PE-based incremental retraining. AL selects data (here translation inputs and their MT outputs for PE) according to criteria that maximise performance with respect to the remaining data and may diverge from processing data items in chronological order. This may allow incrementally PE-based retrained MT to (i) improve more rapidly than chronologically PE-based retrained MT and (ii) result in overall productivity gains.

The main contributions of this paper include:

- Previous work (Haffari et al., 2009; Bloodgood and Callison-Burch, 2010) shows that, given a (static) training set, AL can improve the quality of MT. By contrast, here we show that AL-based data selection for human PE improves incrementally and dynamically retrained MT, reducing overall PE time of translation jobs in the localisation industry application scenarios.
- We propose novel selection criteria for AL-based PE: we adapt cross-entropy difference (Moore and Lewis, 2010; Axelrod et al., 2011), originally used for domain adaptation, and propose an extension to cross entropy difference with a vocabulary saturation filter (Lewis and Eetemadi, 2013).
- While much of previous work concentrates on research datasets (e.g. Europarl, News Commentary), we use industry data (techni-

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 185–189,
Gothenburg, Sweden, April 26-30 2014.

cal manuals). (Bertoldi et al., 2013) shows that the repetition rate of news is considerably lower than that of technical documentation, which impacts on the results obtained with incremental retraining.

- Unlike in previous research, our AL-based selection criteria take into account only the source side of the data. This supports selection before translation, keeping costs to a minimum, a priority in commercial PE MT applications.
- Our experiments show that AL-based selection works for PE-based incrementally retrained MT with overall performance gains around 0.67 to 2.65 TER absolute on average.

AL has been successfully applied to many tasks in natural language processing, including parsing (Tang et al., 2002), named entity recognition (Miller et al., 2004), to mention just a few. See (Olsson, 2009) for a comprehensie overview of the application of AL to natural language processing. (Haffari et al., 2009; Bloodgood and Callison-Burch, 2010) apply AL to MT where the aim is to build an optimal MT model from a given, static dataset. To the best of our knowledge, the most relevant previous research is (González-Rubio et al., 2012), which applies AL to interactive MT. In addition to differences in the AL selection criteria and data sets, our goals are fundamentally different: while the previous work aimed at reducing human effort in interactive MT, we aim at reducing the overall PE time in PE-based incremental MT update applications in the localisation industry.

In our experiments reported in Section 3 below we want to explore a space consisting of a considerable number of selection strategies and incremental retraining batch sizes. In order to be able to do this, we use the target side of our industry translation memory data to approximate human PE output and automatic TER (Snover et al., 2006) scores as a proxy for human PE times (O'Brien, 2011).

2 Methodology

Given a translation job, our goal is to reduce the overall PE time. At each stage, we select sentences that are given to the post editor in such a way that uncertain sentences (with respect to the MT system at hand)[1] are post-edited first. We then translate the n top-ranked sentences using the MT system and use the human PEs of the MT outputs to retrain the system. Algorithm 1 describes our method, where s and t stand for source and target, respectively.

Algorithm 1 Sentence Selection Algorithm

```
Input:
L ⟵ Initial training data
M ⟵ Initial MT model
for C ∈ (Random,Sequential,Ngram,CED,CEDN) do
   U ⟵ Translation job
   while size(U) > 0 do
      U1.s ⟵ SelectTopSentences(C, U.s)
      U1[1].t ⟵ Translate(M, U1.s)
      U1.t ⟵ PostEdit(U1[1].t)
      U ⟵ U - U1
      L ⟵ L ∪ U1
      M ⟵ TrainModel (L)
   end while
end for
```

We use two baselines, i.e. random and sequential. In the random baseline, the batch of sentences at each iteration are selected randomly. In the sequential baseline, the batches of sentences follow the same order as the data.

Aside from the *Random* and *Sequential* baselines we use the following selection criteria:

- **N-gram Overlap.** An SMT system will encounter problems translating sentences containing n-grams not seen in the training data. Thus, PEs of sentences with high number of unseen n-grams are considered to be more informative for updating the current MT system. However, for the MT system to translate unseen n-grams accurately, they need to be seen a minimum number V times.[2] We use an n-gram overlap function similar to the one described in (González-Rubio et al., 2012) given in Equation 1 where $N(T^{(i)})$ and $N(S^{(i)})$ return i-grams in training data and the sentence S, respectively.

$$unseen(S) = \frac{\sum_{i=1}^{n} \{|N(T^{(i)}) \cap N(S^{(i)})| > V\}}{\sum_{i=1}^{n} N(S^{(i)})} \quad (1)$$

- **Cross Entropy Difference (CED).** This metric is originally used in data selection (Moore and Lewis, 2010; Axelrod et al., 2011). Given an in-domain corpus I and a general corpus O, language models are built from both,[3] and each sentence in O is scored according to the entropy H difference (Equation

[1] The uncertainty of a sentence with respect to the model can be measured according to different criteria, e.g. percentage of unknown n-grams, perplexity etc.

[2] Following (González-Rubio et al., 2012) we use $V = 10$.

[3] In order to make the LMs comparable they have the same size. As commonly the size of O is considerable bigger than I, this means that the LM for O is built from a subset of the same size as I.

2). The lower the score given to a sentence, the more useful it is to train a system for the specific domain I.

$$score(s) = H_I(s) - H_O(s) \quad (2)$$

In our AL scenario, we have the current training corpus *L* and an untranslated corpus *U*. CED is applied to select sentences from *U* that are (i) different from *L* (as we would like to add sentences that add new information to the model) and (ii) similar to the overall corpus *U* (as we would like to add sentences that are common in the untranslated data). Hence we apply CED and select sentences from *U* that have high entropy with respect to *L* and low entropy with respect to *U* (Equation 3).

$$score(s) = H_U(s) - H_L(s) \quad (3)$$

- **CED + n-gram (CEDN).** This is an extension of the CED criterion inspired by the concept of the vocabulary saturation filter (Lewis and Eetemadi, 2013). CED may select many very similar sentences, and thus it may be the case that some of them are redundant. By post-processing the selection made by CED with vocabulary saturation we aim to spot and remove redudant sentences. This works in two steps. In the first step, all the sentences from *U* are scored using the CED metric. In the second step, we down-rank sentences that are considered redundant. The top sentence is selected, and its n-grams are stored in *localngrams*. For the remaining sentences, one by one, their n-grams are matched against *localngrams*. If the intersection between them is lower than a predefined threshold, the current sentence is added and *localngrams* is updated with the n-grams from the current sentence. Otherwise the sentence is down-ranked to the bottom. In our experiments, the value n = 1 produces best results.

3 Experiments and Results

We use technical documentation data taken from Symantec translation memories for the English–French (EN–FR) and English–German (EN–DE) language pairs (both directions) for our experiments. The statistics of the data (training and incremental splits) are shown in Table 1.

All the systems are trained using the Moses (Koehn et al., 2007) phrase-based statistical MT system, with IRSTLM (Federico et al., 2008) for language modelling (n-grams up to order five) and with the alignment heuristic *grow-diag-final-and*.

For the experiments, we considered two settings for each language pair in each direction. In the first setting, the initial MT system is trained using the training set (39,679 and 54,907 sentence pairs for EN–FR and EN–DE, respectively). Then, a batch of 500 source sentences is selected from the incremental dataset according to each of the selection criteria, and translations are obtained with the initial MT system. These translations are post-edited and the corrected translations are added to the training data.[4] We then train a new MT system using the updated training data (initial training data plus PEs of the first batch of sentences). The updated model will be used to translate the next batch. The same process is repeated until the incremental dataset is finished (16 and 20 iterations for English–French and English–German, respectively). For each batch we compute the TER score between the MT output and the refererence translations for the sentences of that batch. We then compute the average TER score for all the batches. These average scores, for each selection criterion, are reported in Table 2.

In the second setting, instead of using the whole training data, we used a subset of (randomly selected) 5,000 sentence pairs for training the initial MT system and a subset of 20,000 sentences from the remaining data as the incremental dataset. Here we take batches of 1,000 sentences (thus 20 batches). The results are shown in Table 3.

The first setting aims to reflect the situation where a translation job is to be completed for a domain for which we have a considerable amount of data available. Conversely, the second setting reflects the situation where a translation job is to be carried out for a domain with little (if any) available data.

Dir	Random	Seq.	Ngram	CED	CEDN
EN→FR	29.64	29.81	**28.97**	29.25	29.05
FR→EN	27.08	27.04	**26.15**	26.63	26.39
EN→DE	24.00	24.08	22.34	22.60	**22.32**
DE→EN	19.36	19.34	17.70	17.97	**17.48**

Table 2: TER average scores for Setting 1

Dir	Random	Seq.	Ngram	CED	CEDN
EN→FR	36.23	36.26	35.20	35.48	**35.17**
FR→EN	33.26	33.34	32.26	32.69	**32.17**
EN→DE	32.23	32.19	30.58	31.96	**29.98**
DE→EN	27.24	27.29	26.10	26.73	**24.94**

Table 3: TER average scores for Setting 2

For Setting 1 (Table 2), the best result is obtained by the CEDN criterion for two out of the four directions. For EN→FR, n-gram overlap

[4] As this study simulates the post-editing, we use the references of the translated segments instead of the PEs.

Type	EN–FR			EN–DE		
	Sentences	*Avg. EN SL*	*Avg. FR SL*	*Sentences*	*Avg. EN SL*	*Avg. DE SL*
Training	39,679	13.55	15.28	54,907	12.66	12.90
Incremental	8,000	13.74	15.50	10,000	12.38	12.61

Table 1: Data Statistics for English–French and English–German Symantec Translation Memory Data. SL stands for sentence length, EN stands for English, FR stands for French and DE stands for German

performs slightly better than CEDN (0.08 points lower) with a decrease of 0.67 and 0.84 points when compared to the baselines (random and sequential, respectively). For FR→EN, n-gram overlap results in a decrease of 0.93 and 0.89 points compared to the baselines. The decrease in average TER score is higher for the EN→DE and for DE→EN directions, i.e. 1.68 and 1.88 points respectively for CEDN compared to the random baseline.

In the scenario with limited data available beforehand (Table 3), CEDN is the best performing criterion for all the language directions. For the EN–FR and FR–EN language pairs, CEDN results in a decrease of 1.06 and 1.09 points compared to the random baseline. Again, the decrease is higher for the EN–DE and DE–EN language pairs, i.e. 2.25 and 2.30 absolute points on average.

Figure 1 shows the TER scores per iteration for each of the criteria, for the scenario DE→EN Setting 2 (the trends are similar for the other scenarios). The two baselines exhibit slight improvement over the iterations, both starting at around .35 TER points and finishing at around .25 points. Conversely, all the three criteria start at very high scores (in the range [.5,.6]) and then improve considerably to arrive at scores below .1 for the last iterations. Compared to Ngram and CED, CEDN reaches better scores earlier on, being the criterion with the lowest score up to iteration 13.

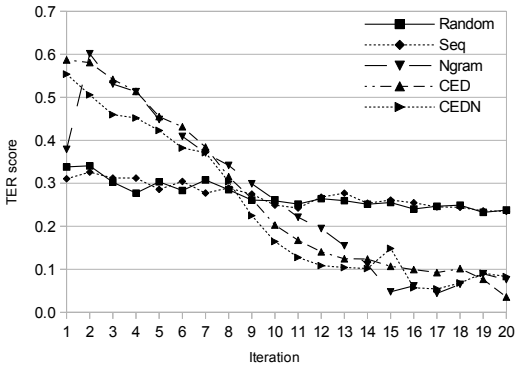

Figure 1: Results per iteration, DE→EN Setting 2

Figure 1 together with Tables 2 and 3 show that AL for PE-based incremental MT retraining really works: all AL based methods (Ngram, CED, CEDN) show strong improvements over both baselines after the initial 8-9 iterations (Figure 1) and best performance on the complete incremental data sets, resulting in a noticeable decrease of the overall TER score (Tables 2 and 3). In six out of eight scenarios, our novel metric CEDN obtains the best result.

4 Conclusions and Future Work

This paper has presented an application of AL to MT for dynamically selecting automatic translations of sentences for human PE, with the aim of reducing overall PE time in a PE-based incremental MT retraining scenario in a typical industrial localisation workflow that aims to capitalise on human PE as early as possible to avoid repeat mistakes.

Our approach makes use of source side information only, uses two novel selection criteria based on cross entropy difference and is tested on industrial data for two language pairs. Our best performing criteria allow the incrementally retrained MT systems to improve their performance earlier and reduce the overall TER score by around one and two absolute points for English–French and English–German, respectively.

In order to be able to explore a space of selection criteria and batch sizes, our experiments simulate PE, in the sense that we use the target reference (instead of PEs) and approximate PE time with TER. Given that TER correlates well with PE time (O'Brien, 2011), we expect AL-based selection of sentences for human PE to lead to overall reduction of PE time. In the future work, we plan to do the experiments using PEs to retrain the system and measuring PE time.

In this work, we have taken batches of sentences (size 500 to 1,000) and do full retraining. As future work, we plan to use fully incremental retraining and perform the selection on a sentence-by-sentence basis (instead of taking batches).

Finally and importantly, a potential drawback of our approach is that by dynamically selecting individual sentences for PE, the human post-editor looses context, which they may use if processing sentences sequentially. We will explore the trade off between the context lost and the productivity gain achieved, and ways of supplying context (e.g. previous and following sentence) for real PE.

Acknowledgements

This work is supported by Science Foundation Ireland (Grants 12/TIDA/I2438, 07/CE/I1142 and 12/CE/I2267) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. We would like to thank Symantec for the provision of data sets used in our experiments.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *Proceedings of the XIV Machine Translation Summit*, pages 35–42, Nice, France.

Michael Bloodgood and Chris Callison-Burch. 2010. Bucking the trend: Large-scale cost-focused active learning for statistical machine translation. In Jan Hajic, Sandra Carberry, and Stephen Clark, editors, *ACL*, pages 854–864. The Association for Computer Linguistics.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.*, 19(2):263–311, June.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *INTERSPEECH*, pages 1618–1621. ISCA.

Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2012. Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 245–254, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *HLT-NAACL*, pages 415–423. The Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej 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*, ACL 2007, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Abby Levenberg, Chris Callison-Burch, and Miles Osborne. 2010. Stream-based translation models for statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 394–402, Stroudsburg, PA, USA. Association for Computational Linguistics.

William Lewis and Sauleh Eetemadi. 2013. Dramatically reducing training data size through vocabulary saturation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 281–291, Sofia, Bulgaria, August. Association for Computational Linguistics.

Prashant Mathur, Mauro Cettolo, and Marcello Federico. 2013. Online learning approaches in computer assisted translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation, ACL*, pages 301–308, Sofia, Bulgaria.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proceedings of HLT*, pages 337–342.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sharon O'Brien. 2011. Towards predicting post-editing productivity. *Machine Translation*, 25(3):197–215, September.

Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing. Technical Report T2009:06.

Mirko Plitt and François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague Bull. Math. Linguistics*, 93:7–16.

Michel Simard and George Foster. 2013. Pepr: Post-edit propagation using phrase-based statistical machine translation. In *Proceedings of the XIV Machine Translation Summit*, pages 191–198, Nice, France.

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*, pages 223–231, Cambridge, MA.

Min Tang, Xiaoqiang Luo, and Salim Roukos. 2002. Active learning for statistical natural language parsing. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 120–127, Stroudsburg, PA, USA. Association for Computational Linguistics.