Online Automatic Post-Editing across Domains

Rajen Chatterjee, Gebremedhen Gebremelak, Matteo Negri, Marco Turchi
Fondazione Bruno Kessler, Italy
{chatterjee, gebremelak, negri, turchi}@fbk.eu

Abstract

English. Recent advances in automatic post-editing (APE) have shown that it is possible to automatically correct systematic errors made by machine translation systems. However, most of the current APE techniques have only been tested in controlled batch environments, where training and test data are sampled from the same distribution and the training set is fully available. In this paper, we propose an online APE system based on an instance selection mechanism that is able to efficiently work with a stream of data points belonging to different domains. Our results on a mix of two datasets show that our system is able to: *i)* outperform state-of-the-art online APE solutions and *ii)* significantly improve the quality of rough MT output.

Italiano. *Recenti miglioramenti dei sistemi automatici di post-editing hanno dimostrato la loro capacità di correggere errori ricorrenti commessi dalla traduzione automatica. Spesso, tuttavia, tali sistemi sono stati valutati in condizioni controllate dove i dati di training/test sono selezionati dalla stessa distribuzione e l'insieme di training è interamente disponibile. Questo articolo propone un sistema di post-editing online, basato su tecniche di selezione dei dati, capace di trattare sequenze di dati appartenenti a diversi dominii. I risultati su un insieme di dati misti mostrano che il sistema è in grado di ottenere risultati migliori rispetto i) allo stato dell'arte e ii) al sistema di traduzione.*

1 Introduction

Nowadays, machine translation (MT) is a core element in the computer-assisted translation (CAT) framework (Federico et al., 2014). The motivation for integrating MT in the CAT framework lies in its capability to provide useful suggestions for unseen segments, thus increasing translators productivity. However, it has been observed that MT is often prone to systematic errors that human post-editing has to correct before publication. The by-product of this "translation as post-editing" process is an increasing amount of parallel data consisting of MT output on one side and its corrected version on the other side. Besides being used to improve the MT system itself (Bentivogli et al., 2016), this data can be leveraged to develop automatic MT quality estimation tools (Mehdad et al., 2012; Turchi et al., 2013; C. de Souza et al., 2013; C. de Souza et al., 2014; C. de Souza et al., 2015) and automatic post-editing (APE) systems (Chatterjee et al., 2015b; Chatterjee et al., 2015a; Chatterjee et al., 2016). The APE components explored in this paper should be capable not only to spot recurring MT errors, but also to correct them. Thus, integrating an APE system inside the CAT framework can further improve the quality of the suggested segments, reduce the workload of human post-editors and increase the productivity of translation industries. In the last decade many studies on APE have shown that the quality of the machine translated text can be improved significantly by post-processing the translations with an APE system (Simard et al., 2007; Dugast et al., 2007; Terumasa, 2007; Pilevar, 2011; Béchara et al., 2011; Chatterjee et al., 2015b). These systems mainly follow the phrase-based machine translation approach where the MT outputs (with optionally the source sentence) are used as the source language corpus and the post-edits are used as the target language corpus. Although these standard approaches showed promising results, they lack of the ability to continuously update their inner models by incorporating human feedback from a stream of data. To address this problem, several online systems have been proposed in MT, but only few of them have been applied to the APE scenario (Simard and Foster, 2013; Lagarda et al., 2015), only in a controlled working environment where they are trained and evaluated on homogeneous/coherent data sets.

In this paper, we propose a novel online APE system that is able to efficiently leverage data from different domains.[1] Our system is based on an instance selection technique that is able to retrieve the most relevant training instances from a pool of *multi-domain* data for each segment to post-edit. The selected data is then used to train and tune the APE system on-the-fly. The relevance of a training sample is measured by a similarity score that takes into account the context of the segment to be post-edited. This technique allows our online APE system to be flexible enough to decide if it has the correct knowledge for post-editing a sentence or if it is safer to keep the MT output untouched, avoiding possible damages. The results of our experiments over the combination of two data sets show that our approach is robust enough to work in a multi-domain environment and to generate reliable post-edits with significantly better performance than a state-of-the-art online APE system.

2 Online translation systems

Online translation systems aim to incorporate human post-editing feedback (or the corrected version of the MT output) into their models in real-time, as soon as it becomes available. This feedback helps the system to learn from the mistakes made in the past translations and avoid to repeat them in future translations. This continuous learning capability will eventually improve the quality of the translations and consequently increase the productivity of the translators/post-editors (Tatsumi, 2009) working with MT suggestions in a CAT environment. The basic workflow of an online translation system goes through the following steps repeatedly: *i)* the system receives an input segment; *ii)* the input segment is translated and provided to the post-editor to fix any errors in it; and *iii)* the human post-edited version of the translation is incorporated back into the system, by stepwise updating the underlying models and parameters. In the APE context, the input is a machine-translated segment (optionally with its corresponding source segment), which is processed by the online APE system to fix errors, and then verified by the post-editors. Several online translation systems have been proposed over the years (Hardt and Elming, 2010; Bertoldi et al., 2013; Mathur et al., 2013; Simard and Foster, 2013; Ortiz-Martı̈nez and Casacuberta, 2014; Denkowski et al., 2014; Wuebker et al., 2015).

The state-of-the-art online APE system is the Thot toolkit (Ortiz-Martı̈nez and Casacuberta, 2014) that has been previously developed to support fully automatic and interactive statistical machine translation and then used in the APE task (Lagarda et al., 2015). To update the inner models with the user feedback, a set of sufficient statistics was maintained and incrementally updated. In the case of language model, only the n-gram counts are required to maintain sufficient statistics. To update the translation model, an incremental version of EM algorithm is used to first obtain word alignment and then phrase pairs counts were extracted to update the sufficient statistics. Other features like source/target phrase-length models or distortion model are implemented by means of geometric distributions with fixed parameters. However, Thot differs from our approach because it does not embed any techniques for selecting the most relevant training data. In the long-run, when data points from different domain are continuously analysed, this system tends to become more and more generic, which may not be useful and even harmful for automatically post-editing domain-specific segments.

3 Instance Selection for online APE system

To preserve all the knowledge gained in the online learning process and at the same time being able to apply specific post-editing rules when needed, we propose an instance selection technique for online APE that has the ability to retrieve specific data points whose context is similar to the segment to be post-edited. These data points are then used to build reliable APE models. When there are no reliable data points in the knowledge base, the MT output is kept untouched, as opposed to the exist-

[1] A domain is made of segments belonging to the same text genre and the MT outputs are generated by the same MT system.

ing APE systems, which tends to always translate the given input segment independently from the reliability of the applicable correction rules.

Our proposed algorithm emulates an online APE system and assumes to have the following data to run the online experiments: *i)* source (*src*); *ii)* MT output (*mt*); and *iii)* human post-edits (*pe*) of the MT output. At the beginning the knowledge base of our online APE system is empty and it will be updated whenever an instance (a tuple containing parallel segments from all the above mentioned documents) is processed. When the system receives an input (*src*, *mt*), the most relevant training instances from a pool of *multi-domain* data stored in our knowledge base are retrieved. The similarity between the training instances and the input segment is measured by a score based on the term frequency−inverse document frequency (*tf-idf*), generally used in information retrieval. The larger the number of words in common between the training and the input sentences, the higher is the score. In our system, these scores are computed using the Lucene library.[2] Only those training instances that have similarity score above a certain threshold (decided over a held-out development set) are used to build: *i)* a tri-gram local language model over the target side of the training corpus with the IRSTLM toolkit (Federico et al., 2008); ii) the translation and reordering models using the Moses toolkit (Koehn et al., 2007) and the word alignment of each sentence pair is computed using the incremental GIZA++ software.[3] The log-linear model parameters are optimized over a part of the selected instances. To obtain reliably-tuned weights and a fast optimization process, multiple instances of MIRA (Chiang, 2012) are run in parallel on three small development sets randomly selected from the retrieved sentences. The obtained weights are then averaged. If a minimum value of retrieved sentences is not reached, the optimization step is skipped because having few sentences might not yield reliable weights. In this case, the weights computed on the previous input segment are used. The tuned weights and the models built on all the data are then used to post-edit the input sentences.

In a real translation workflow, the APE segment is then passed to the human translator that creates the post-edited segment. Once the post-edit is available it is added to the knowledge base along with the source and the mt sentences. In our experiments we emulate the post-edited sentence of the APE segment with the post-edit of the mt output.

4 Experimental setup

Data To examine the performance of the online APE systems in a *multi-domain* translation environment, we select two data sets for the English-German language pair belonging to information technology (IT). Although they come from the same category (IT), they feature variability in terms of vocabulary coverage, MT errors, and post-editing style. The two data sets are respectively a subset of the Autodesk Post-Editing Data corpus and the resources used at the second round of the APE shared task at the first conference on machine translation (WMT2016).[4] The data sets are pre-processed to obtain a joint-representation that links each source word with a MT word (*mt#src*). This representation has been proposed in the context-aware APE approach by (Béchara et al., 2011) and leverages the source information to disambiguate post-editing rules. Recently, (Chatterjee et al., 2015b) also confirmed this approach to work better than translating from raw MT segments over multiple language pairs. The joint-representation is used as a source corpus to train all the APE systems reported in this paper and it is obtained by first aligning the words of source (*src*) and MT (*mt*) segments using MGIZA++ (Gao and Vogel, 2008), and then each *mt* word is concatenated with its corresponding *src* words.

The Autodesk training, and development sets consist of 12,238, and 1,948 segments respectively, while the WMT2016 data contains 12,000, and 1,000 segments. To measure the diversity of the two data sets we compute the vocabulary overlap between the two joint-representations. This is performed internally to each data set (splitting the training data in two halves) and across them. As expected, in the first case the vocabulary overlap is much larger ($> 40\%$) than in the second one ($\sim 15\%$); this indicates that the two data sets are quite different and few information can be shared.

To emulate the multi-domain scenario, the two training data sets are first merged together and then shuffled. The same strategy is also used for the development sets. This represents the situation in

[2] https://lucene.apache.org/

[3] https://code.google.com/archive/p/inc-giza-pp/

[4] http://www.statmt.org/wmt16/ape-task.html

which an APE system serves two CAT tools that process documents from two domains and the sequence of points is random. Our approach and the competitors are run on all the shuffled training data and evaluated on the second half (12,100 points).

Evaluation metrics The performance of the different APE systems is evaluated using the Translation Error rate (TER) (Snover et al., 2006), BLEU (Papineni et al., 2002) and the precision (Chatterjee et al., 2015a). TER and BLEU measures the similarity between the MT outputs and their references by looking at the word/n-gram overlaps, while precision is the ratio of number of sentences an APE system improves (with respect to the MT output) over all the sentences it modifies.[5] Larger values indicate that the APE system is able to improve the quality of most of the sentences it changes. The statistical significance test for BLEU is computed using the paired bootstrap resampling technique (Koehn, 2004), and for TER using the stratified approximate randomization technique (Clark et al., 2011).

Terms of comparison We evaluate our online learning approach against the output produced by the *MT system*, the *batch APE system* that follows the approach proposed in (Chatterjee et al., 2015b), and the Thot toolkit.

5 Experiments and Results

The main goal of this research is to examine the performance of online APE methods in a multi-domain scenario, where the APE system receives a stream of data coming from different domains. The parameters of our approach (*i.e.* similarity score threshold and minimum number of selected sentence) are optimised following the grid search strategy. We set the threshold values to 1 and the minimum number of selected sentences to 20. The results of all the systems are reported in Table 1.

The *batch APE system* that is trained only on the first half of the data is able to slightly improve the performance of the *MT system*, but it damages most of the sentence it changes (precision smaller than 45%). Although Thot can learn from all the data, it is interesting to note that it does not significantly improve over the *MT system* and the *batch APE system*. This suggests that using all the data without considering the peculiarities of each domain does not allow an APE system to efficiently learn reliable correction rules and to improve the machine translation quality. Moreover, these results also show that few information can be shared between the two data sets. This is expected considering the limited overlap between the two corpora.

	BLEU	TER	Precision (%)
MT	52.31	34.52	N/A
Batch APE	52.52	34.45	42.67
Thot	52.51	34.37	42.22
Our approach	**53.97**†	**33.13**†	**64.82**

Table 1: Results on the mixed data. (†: statistically significant wrt. MT with p<0.05)

Our approach provides significant improvements in BLEU, TER and precision over all the competitors. In particular, it can obtain more than one TER and BLEU point improvement, and more than 20% precision points increment over the best APE system (the Thot toolkit). Such gains confirm that the instance selection mechanism allows our APE system to identify domain-specific data and to leverage it for extracting reliable correction rules. Further analysis of the performance of the online systems revealed that our approach modifies less segments compared with Thot, because it builds a model only if it finds relevant data, leaving the MT segment untouched otherwise. These untouched MT segments, when modified by Thot, often lead to deterioration. This suggests that, the output obtained with our solution has a higher potential for being useful to human translators. Such usefulness comes not only in terms of a more pleasant post-editing activity, but also in terms of time savings yield by overall better suggestions.

6 Conclusion

We addressed the problem of building a robust online APE system that is able to efficiently work on a stream of data points belonging to different domains. In this condition, our APE has shown its capability to continuously adapt to the dynamics of diverse data processed in real-time. In particular, the instance selection mechanism allows our APE method to reduce the number of wrong modifications, which result in significant improvements in precision over the state-of-the-art online APE system, and thus making it a viable solution to be deployed in a real-word CAT framework.

[5] For each sentence in the test set, if the TER score of APE system is different than the baseline then it is considered as a modified sentence

7 Acknowledgements

This work has been partially supported by the EC-funded H2020 project QT21 (grant agreement no. 645452).

References

Hanna Béchara, Yanjun Ma, and Josef van Genabith. 2011. Statistical post-editing for a statistical mt system. In *Proceedings of the XIII MT Summit*, pages 308–315.

Luisa Bentivogli, Nicola Bertoldi, Mauro Cettolo, Marcello Federico, Matteo Negri, and Marco Turchi. 2016. On the evaluation of adaptive machine translation for human post-editing. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 24(2):388–399.

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

José G. C. de Souza, Christian Buck, Marco Turchi, and Matteo Negri. 2013. FBK-UEdin Participation to the WMT13 Quality Estimation Shared Task. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, pages 352–358, Sofia, Bulgaria. Association for Computational Linguistics.

José G. C. de Souza, Jesús González-Rubio, Christian Buck, Marco Turchi, and Matteo Negri. 2014. FBK-UPV-UEdin Participation in the WMT14 Quality Estimation Shared-task. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 322–328, Baltimore, Maryland, USA.

José G. C. de Souza, Matteo Negri, Elisa Ricci, and Marco Turchi. 2015. Online multitask learning for machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 219–228, Beijing, China, July. Association for Computational Linguistics.

Rajen Chatterjee, Marco Turchi, and Matteo Negri. 2015a. The fbk participation in the wmt15 automatic post-editing shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 210–215.

Rajen Chatterjee, Marion Weller, Matteo Negri, and Marco Turchi. 2015b. Exploring the planet of the apes: a comparative study of state-of-the-art methods for mt automatic post-editing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 156–161, July.

Rajen Chatterjee, José G. C. de Souza, Matteo Negri, and Marco Turchi. 2016. The fbk participation in the wmt 2016 automatic post-editing shared task. In *Proceedings of the First Conference on Machine Translation*, pages 745–750, Berlin, Germany, August. Association for Computational Linguistics.

David Chiang. 2012. Hope and fear for discriminative training of statistical translation models. *Journal of Machine Learning Research*, 13(Apr):1159–1187.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 176–181.

Michael Denkowski, Chris Dyer, and Alon Lavie. 2014. Learning from post-editing: Online model adaptation for statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 395–404, April.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical post-editing on systran's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. Irstlm: an open source toolkit for handling large scale language models. In *Proceedings of Interspeech*, pages 1618–1621.

Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, Alberto Massidda, Holger Schwenk, Loïc Barrault, Frederic Blain, Philipp Koehn, Christian Buck, and Ulrich Germann. 2014. The matecat tool. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 129–132, Dublin, Ireland, August.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Proceedings of Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Daniel Hardt and Jakob Elming. 2010. Incremental re-training for post-editing smt. In *Proceedings of AMTA*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pages 177–180.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, pages 388–395.

Antonio L Lagarda, Daniel Ortiz-Martïnez, Vicent Alabau, and Francisco Casacuberta. 2015. Translating without in-domain corpus: Machine translation post-editing with online learning techniques. *Computer Speech & Language*, 32(1):109–134.

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

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012. Match without a Referee: Evaluating MT Adequacy without Reference Translations. In *Proceedings of the Machine Translation Workshop (WMT2012)*, pages 171–180, Montréal, Canada, June.

Daniel Ortiz-Martïnez and Francisco Casacuberta. 2014. The new thot toolkit for fully-automatic and interactive statistical machine translation. In *14th Annual Meeting of the European Association for Computational Linguistics: System Demonstrations*, pages 45–48.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Abdol Hamid Pilevar. 2011. Using statistical post-editing to improve the output of rule-based machine translation system. *IJCSC*.

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

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical Phrase-Based Post-Editing. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 508–515.

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 AMTA*, pages 223–231.

Midori Tatsumi. 2009. Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. In *Proceedings of the XII MT Summit*, pages 332–339.

Ehara Terumasa. 2007. Rule based machine translation combined with statistical post editor for japanese to english patent translation. In *Proceedings of the XI MT Summit*, pages 13–18.

Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the Subjectivity of Human Judgements in MT Quality Estimation. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria. Association for Computational Linguistics.

Joern Wuebker, Spence Green, and John DeNero. 2015. Hierarchical incremental adaptation for statistical machine translation. In *Proceedings of EMNLP*, pages 1059–1065, September.