Exploring the Planet of the APEs: a Comparative Study of State-of-the-art Methods for MT Automatic Post-Editing

Rajen Chatterjee[1,2], **Marion Weller**[3], **Matteo Negri**[2], **Marco Turchi**[2]
[1] University of Trento
[2] FBK - Fondazione Bruno Kessler
[3] IMS, University of Stuttgart
{chatterjee,negri,turchi}@fbk.eu
{wellermn@ims.uni-stuttgart.de}

Abstract

Downstream processing of machine translation (MT) output promises to be a solution to improve translation quality, especially when the MT system's internal decoding process is not accessible. Both rule-based and statistical automatic post-editing (APE) methods have been proposed over the years, but with contrasting results. A missing aspect in previous evaluations is the assessment of different methods: *i)* under comparable conditions, and *ii)* on different language pairs featuring variable levels of MT quality. Focusing on statistical APE methods (more portable across languages), we propose the first systematic analysis of two approaches. To understand their potential, we compare them in the same conditions over six language pairs having English as source. Our results evidence consistent improvements on all language pairs, a relation between the extent of the gain and MT output quality, slight but statistically significant performance differences between the two methods, and their possible complementarity.

1 Introduction

Automatic post-editing (APE) aims to correct systematic machine translation (MT) errors. The problem is appealing for several reasons. On one side, as pointed out by Parton et al. (2012), APE systems can improve MT output by exploiting information unavailable to the decoder, or by performing deeper text analysis that is too expensive at decoding stage. On the other side, and to our view more importantly, APE represents the only way to recover errors produced in "black-box" conditions in which the MT system is unknown or its internal decoding process is not accessible.

The task, firstly proposed by Knight and Chander (1994) to cope with article selection in Japanese to English translation, has been later addressed in various ways. On one side, rule-based methods (Rosa et al., 2012) gained limited attention, probably due to the extensive manual work they involve and their scarce portability across languages. On the other side, the statistical approach proposed by Allen and Hogan (2000) reached maturity in the work by Simard et al. (2007) and inspired a number of further investigations (Isabelle et al., 2007; Dugast et al., 2007; Dugast et al., 2009; Lagarda et al., 2009; Béchara et al., 2011; Béchara et al., 2012; Rubino et al., 2012; Rosa et al., 2013; Lagarda et al., 2014, inter alia).

Such prior works address orthogonal aspects like: *i)* performance variations when APE is applied to correct the output of *rule-based vs. statistical MT*, *ii)* the use of APE for *error correction vs. domain adaptation*, *iii)* the difference between training on *general domain vs. domain-specific data*, *iv)* performance variations when learning from *reference translations vs. human post-edits*. Their common trait is that the reported results are difficult to generalise. Indeed, most of the works focus on evaluating a specific method,[1] which is typically applied to one single dataset for a given language pair. As a result, the global landscape of the "planet of the APEs" is still blurred and open to more systematic explorations.

To shed light on the potential of statistical post-editing, in this paper we examine two alternative approaches. One is the method proposed in (Simard et al., 2007), which to date is the most widely used. The other is the "context-aware" solution proposed in (Béchara et al., 2011) which, to the best of our knowledge, represents the most significant variant of (Simard et al., 2007).

The major contribution of our work is the first systematic analysis of different APE approaches,

[1] Typically the same of (Simard et al., 2007).

which are tested in controlled conditions over several language pairs. To ensure the soundness of the analysis, our experimental setup consists of a dataset composed of the same English source sentences with automatic translations into six languages and respective manual post-edits by professional translators. Overall, this represents the ideal condition to complement prior research with the missing answers to questions like:

Q1: *Does APE yield consistent MT quality improvements across different language pairs?*

Q2: *What is the relation between the original MT output quality and the APE results?*

Q3: *Which of the two analysed APE methods has the highest potential?*

2 Statistical APE methods

The two methods we analyse follow the same "statistical phrase-based post-editing" strategy outlined by Simard et al. (2007), but differ in the way data is represented. Let's give them a closer look.

2.1 Method 1 (Simard et al., 2007)

The underlying idea is that APE components can be trained in the same way in which statistical MT systems are developed – *i.e.* starting from "parallel data". Since the goal is to transform rough MT output into its correct version, parallel data consists of MT output as source texts and correct (human quality) sentences as target. In (Simard et al., 2007) these are used to train a phrase-based MT system, which is then applied to correct the output of a commercial rule-based MT system.

Positive evaluation results are reported on English-French, and even better ones on French-English data. In both cases, statistical APE yields significant BLEU and TER improvements over the original MT output. However, since training and test data for the two language directions are different (in content and size), the measured performance variations cannot be directly ascribed to the effectiveness of the method in the two settings.

2.2 Method 2 (Béchara et al., 2011)

One limitation of the "monolingual translation" approach proposed in (Simard et al., 2007) is that the basic statistical APE pipeline is only trained on data in the target language (F), disregarding information about the source language (E): Correction rules learned from (f', f) pairs[2] lose the connection between the translated words (or phrases) and the corresponding source terms (e). This implies that information lost or distorted in the translation process is out of the reach of the APE component, and the resulting errors are impossible to recover.

To cope with this issue, Béchara et al. (2011) propose a "context-aware" variant to represent the data. For each word f', the corresponding source word (or phrase) e is identified through word alignment and used to obtain a joint representation $f'\#e$. The result is an intermediate language $F'\#E$ that represents the new source side of the parallel data used to train the statistical APE component. Though in principle more precise, this method can be affected by two problems. First, preserving the source context comes at the cost of a larger vocabulary size and, consequently, higher data sparseness. While the basic statistical APE pipeline combines and exploits the counts of all the co-occurrences of f' and f in the parallel data, its context-aware variant considers each $f'\#e_i$ as a separate term, thus breaking down the co-occurrence counts of f' and f into smaller numbers. Second, all these counts can be influenced by word alignment errors. To cope with data sparseness and unreliable word alignment, Béchara et al. (2011) experiment with different thresholds set on word alignment strengths to filter context information. In particular, they discard the $(f'\#e, f)$ pairs in which the $f'\#e$ alignment score is smaller than the threshold.

The approach, applied to correct the output of a statistical phrase-based MT system, achieves ambiguous evaluation results. On French-English, significant improvements up to 2 BLEU points are observed both over the baseline (the original MT output) and the basic method of Simard et al. (2007). On English-French, however, performance slightly drops. Moreover, follow-up experiments with the same method (Béchara, 2014) did not confirm these results. *In light of these ambiguous results and the lack of a systematic comparison between the two APE methods, our objective is to replicate them*[3] *for a fair comparison in a controlled evaluation setting involving different lan-*

[2] Here, f' and f respectively stand for the rough MT output and its correct version in the foreign language F.

[3] This is done based on the description provided by the published works. Discrepancies with the actual methods are possible, due to our misinterpretation or to wrong guesses about details that are missing in the papers.

guage pairs.

2.3 Reimplementing the two methods

To obtain the statistical APE pipeline that represents the backbone of both methods we used a phrase-based Moses system (Koehn et al., 2007). Our training data (see Section 3) consists of (*source*, *MT output*, *post-edition*) triplets for six language pairs having English as source. While Method 1 uses only the last two elements of the triplet, all of them play a role in the context-aware Method 2. Apart from the different data representation, the training process is identical.

Translation and reordering models were estimated following the Moses protocol with default setup using MGIZA++ (Gao and Vogel, 2008) for word alignment.[4] For language modeling we used the KenLM toolkit (Heafield, 2011) for standard n-gram modeling with an n-gram length of 5. The APE system for each target language was tuned on comparable development sets (see below), optimizing TER with Minimum Error Rate Training (Och, 2003) using the post-edited sentences as references.

3 Experiments

Some lessons learned from prior works on statistical APE methods (Béchara, 2014) include: *i)* learning from human post-edits is more effective than learning from (independent) reference translations, *ii)* learning from (and applying APE to) domain-specific data is more promising than working on general-domain data, *iii)* correcting the output of rule-based MT systems is easier than improving translations from statistical MT. Our work capitalizes on these findings (we learn from domain-specific post-edited data and apply APE to statistical MT), but fills a gap of previous research: a fair comparative study between different methods in controlled conditions. The key enabling factor is the availability, for the first time, of data consisting of the *same source sentences*, *machine-translated in several languages* and *post-edited by professional translators*.

Data. We experiment with the Autodesk Post-Editing Data corpus,[5] which predominantly covers the domain of software user manuals. English sentences are translated into several languages (30K to 410K translations per language) with Autodesk's in-house MT system (Zhechev, 2012) and post-edited by professional translators.

Lang.	No. tokens	Vocab. Size	No. Lemmas
En	210,491	10,727	8,260
Cs	202,475	16,716	10,137
De	211,149	17,563	14,368
Es	252,020	11,075	6,683
Fr	263,690	10,928	7,213
It	239,912	10,703	6,549
Pl	206,016	17,027	10,430

Table 1: Data statistics for each language.

Our experiments are run on six language pairs having English as source and Czech, German, Spanish, French, Italian and Polish as target. To set up our controlled environment, we extract all the (*source*, *MT output*, *post-edition*) triplets sharing the same source (En) sentences across all language pairs. Table 1 provides some statistics about the resulting *tri-parallel* corpora. After random shuffling the triplets, we create training (12.2K triplets), development (2K) and test data (2K) sharing exactly the same source sentences across languages. Training and evaluation of our APE systems are performed on true-case data.

To guarantee similar experimental conditions in the six language settings, we also train comparable target language models from external data (indeed, the 12.2K post-edits would not be enough to train reliable LMs). We build our LMs from approximately 2.5M translations of the same English sentences collected from Europarl (Koehn, 2005), DGT-Translation Memory (Steinberger et al., 2012), JRC Acquis (Steinberger et al., 2006), OPUS IT (Tiedemann,) and other Autodesk data common to all languages.

Evaluation metric. We evaluate the APE methods based on their capability to reduce the distance between the MT output and a correct (fluent and adequate) translation. As a measure of the amount of the editing operations needed for the correction, TER and HTER (Snover et al., 2006) fit for our purpose. TER and HTER measure the minimum edit distance between the MT output and its cor-

[4] In Method 1, MGIZA++ is used to align f' and f. In Method 2 it is used to align f' and e, and then $f'\#e$ and f.

[5] `https://autodesk.app.box.com/Autodesk-PostEditing`

	MT Baseline	Method 1			Method 2			Oracle
	TER	*TER*	Δ	*% Reduction*	*TER*	Δ	*% Reduction*	*TER*
En-De	46.46	43.07	-3.39	7.3	42.79*	-3.67	7.9	40.17
En-Cs	44.06	39.38	-4.68	10.62	39.10*	-4.96	11.25	36.32
En-Pl	43.02	38.24	-4.78	11.11	37.75*	-5.27	12.25	35.05
En-It	34.44	30.43	-4.01	11.64	30.13*	-4.31	12.55	28.33
En-Fr	32.76	29.70	-3.06	9.34	29.51	-3.25	9.92	27.12
En-Es	30.90	26.69	-4.21	13.62	26.35*	-4.55	14.72	24.34

Table 2: Performance of the MT baseline and the APE methods for each language pair. Results for Method 2 marked with the "*" symbol are statistically significant compared to Method 1.

rect version.[6] This can be either a reference translation created independently from the MT output (TER) or a human post-edition obtained by manually correcting the MT output (HTER). For the sake of simplicity, henceforth we will use the term TER to refer to both situations (though, when measuring the distance between the MT output and its human post-edition the actual metric is the HTER).

Baseline. Similar to all previous works on APE, our baseline is the MT output *as is*. Hence, baseline scores for each language pair correspond to the TER computed between the original MT output (produced by the "black-box" Autodesk in-house system) and the human post-edits.

4 Results

Table 2 lists our results, with language pairs ordered according to the respective baseline TER. The positive answer to **Q1** (*"Does APE yield consistent improvements to MT output?"*) is evident: both APE methods consistently improve MT quality on all language pairs. TER reductions range from 3.06 to 5.27 points. Quality improvements are statistically significant at $p < 0.05$, measured by bootstrap test (Koehn, 2004).

In answer to **Q2** (*"What is the relation between original MT quality and APE results?"*), our controlled experiments evidence for the first time in APE research that the higher the MT quality, the higher is the improvement, *i.e.* percentage of error reduction, yielded by the APE methods. On one side, this interesting result may seem counter-intuitive because a larger room for improvement is expected for sentences of poor quality. On the other side, it reveals that learning from (and correcting) noisy data affected by many errors is particularly difficult for statistical APE methods. This finding is violated by En-Fr, for which a reasonably good MT quality does not induce a gain in performance comparable to language pairs featuring similar MT TER (En-It and En-Es). On further analysis of the data, we notice that all the target languages except French keep a coherent behaviour with respect to the domain-specific English terms, which are always either preserved (It) or translated (other languages). Instead, French shows an alternation between the two conducts. One example is the English word "*workflow*", which appears in the French post-editions both *as is* (21 sentences) and translated into "*flux de travail*" (34 sentences). In contrast, in the other language directions all the occurrences of '*workflow*" are either translated or kept in English. These frequent ambiguities are difficult to manage (especially if the two forms occur a similar number of times in the training data), and might motivate the smaller quality gains observed on En-Fr compared to the other language pairs.

In answer to **Q3** (*"Which method has the highest potential?"*), we observe slight TER reductions when moving from Method 1 to its "context-aware" variant.[7] Although small (from 0.19 to 0.49 TER points), such gains are statistically significant ($p < 0.05$), except for En-Fr ($p < 0.07$). This suggests that linking the MT words to the source terms can help to recover adequacy errors that are out of the reach of Method 1.

To better understand to what extent the two methods behave differently, we calculated the results of an *Oracle* system, similar to the one pro-

[6] Edit distance is calculated as the number of edits (word insertions, deletions, substitutions, and shifts) divided by the number of words in the correct translation. Lower TER/HTER values indicate better MT quality.

[7] Filtering the context information with thresholds between 0.6 and 0.8 leads to the best results for all languages.

posed by Rubino et al. (2012), defined by selecting for each test sentence the best post-edit (lower TER) produced by two approaches. As shown in the last column of Table 2, such an oracle achieves a significant TER reduction (from 1.8 to 2.78 points) for all the language pairs. We interpret such gains as clues of a possible complementarity between the two methods, which is worth to investigate.

As mentioned in Section 2.2, an advantage of Method 1 is its robust estimation of translation parameters. In contrast, by exploiting contextual information from the source, Method 2 is more precise but potentially affected by data sparsity issues due to its highly increased vocabulary. In an attempt to use a less sparse model at the level of word alignment, we trained a SMT system based on the context-aware representation of Method 2 ($f'\#e$), but with word alignment computed on the representation of Method 1 (f'). Applying this method to the three language pairs for which the two original methods achieved the lowest TER reductions (*i.e.* En-De, En-Fr and En-Cs) shows that this simple way to combine Methods 1 and 2 is able to produce a TER decrement of 0.75 (42.04) for En-De, 0.60 (38.50) for En-Cs and 0.53 (28.98) for En-Fr. This seems to validate our intuition about the possible complementarity of Methods 1 and 2, suggesting a promising direction for future work.

5 Conclusions

We explored the "planet of the APEs" in ideal conditions (quantity and quality of data) and with the right equipment (state-of-the-art methods). The data available (the same English sentences, machine-translated in six languages and post-edited by professional translators) allowed us to compare for the first time different approaches in a fair setting (*our first contribution*). The two methods we analysed allowed us to measure consistent improvements on all language pairs (TER reductions from 7.3% to 14.7% – *second contribution*), and to observe interesting relations between the extent of the gain and the original MT output quality (the higher the quality, the higher the gain yield by APE – *third contribution*). This first study represents a good starting point for future quests. A promising direction to explore is the possible complementarity between the two methods and the room for mutual improvement. Now we just have a glimpse of the path (higher oracle results, slight gains with a first combination method – *fourth contribution*), but positive preliminary results confirm its existence.

To encourage the replication of our experiments by other researchers and the reuse of the selected Autodesk data for benchmarking purposes in the same setting, the scripts developed in this work have been publicly released. They can be downloaded from: `https://bitbucket.org/turchmo/apeatfbk/src/master/papers/ACL2015/`.

Acknowledgements

This work has been partially supported by the EC-funded H2020 project QT21 (grant agreement no. 645452). The work of Marion Weller was supported by the FBK-HLT Summer Internship Program 2014. The authors would like to thank Dr. Ventsislav Zhechev for his support with the Autodesk Post-Editing Data corpus.

References

Jeffrey Allen and Christopher Hogan. 2000. Toward the Development of a Post Editing Module for Raw Machine Translation Output: A Controlled Language Perspective. In *Third International Controlled Language Applications Workshop (CLAW-00)*, pages 62–71.

Hanna Béchara, Yanjun Ma, and Josef van Genabith. 2011. Statistical Post-Editing for a Statistical MT System. In *Proceedings of the 13th Machine Translation Summit*, pages 308–315, Xiamen, China, September.

Hanna Béchara, Raphaël Rubino, Yifan He, Yanjun Ma, and Josef Genabith. 2012. An Evaluation of Statistical Post-Editing Systems Applied to RBMT and SMT Systems. In *Proceedings of COLING 2012*, pages 215–230, Mumbai, India.

Hanna Béchara. 2014. Statistical Post-editing and Quality Estimation for Machine Translation Systems. *M.Sc. Thesis, Dublin City University, Dublin.*

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*, StatMT '07, pages 220–223, Stroudsburg, PA, USA.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2009. Statistical Post Editing and Dictionary Extraction: Systran/Edinburgh Submissions for ACL-WMT2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 110–114, Athens, Greece.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Proceedings of the ACL 2008 Software Engineering, Testing, and Quality Assurance Workshop*, pages 49–57, Columbus, Ohio.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

Pierre Isabelle, Cyril Goutte, and Michel Simard. 2007. Domain Adaptation of MT Systems through Automatic Post-editing. In *Proceedings of the Eleventh Machine Translation Summit (MT Summit XI)*, pages 255–261, Copenhagen, Denmark.

Kevin Knight and Ishwar Chander. 1994. Automated Postediting of Documents. In *Proceedings of the 12th National Conference on Artificial Intelligence*, pages 779–784, Seattle, WA, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand, September.

Antonio L. Lagarda, Vicent Alabau, Francisco Casacuberta, Roberto Silva, and Enrique Díaz-de Liaño. 2009. Statistical Post-editing of a Rule-based Machine Translation System. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 217–220.

Antonio L. Lagarda, Daniel Ortiz-Martínez, Vicent Alabau, and Francisco Casacuberta. 2014. Translating without in-domain Corpus: Machine Translation Post-editing with Online Learning Techniques. *Computer Speech & Language*.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Sapporo, Japan.

Kristen Parton, Nizar Habash, Kathleen McKeown, Gonzalo Iglesias, and Adrià de Gispert. 2012. Can Automatic Post-Editing Make MT More Meaningful? In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 111–118, Trento, Italy.

Rudolf Rosa, David Mareček, and Ondřej Dušek. 2012. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 362–368, Montreal, Canada.

Rudolf Rosa, David Marecek, and Ales Tamchyna. 2013. Deepfix: Statistical Post-editing of Statistical Machine Translation Using Deep Syntactic Analysis. In *Proceedings of the ACL 2013 Student Research Workshop*, pages 172–179, Sofia, Bulgaria.

Raphaël Rubino, Stéphane Huet, Fabrice Lefèvre, and Georges Lenarés. 2012. Statistical Post-Editing of Machine Translation for Domain Adaptation. In *Proceedings of the European Association for Machine Translation (EAMT)*, pages 221–228, Trento, Italy.

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 (NAACL HLT)*, pages 508–515, Rochester, New York.

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, Massachusetts, USA.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dniel Varga. 2006. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genoa, Italy.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2012. DGT-TM: A Freely Available Translation Memory in 22 Languages. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Jörg Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey.

Ventsislav Zhechev. 2012. Machine Translation Infrastructure and Post-Editing Performance at Autodesk. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, pages 87–96, San Diego, CA, USA.