Touch-Based Pre-Post-Editing of Machine Translation Output

Benjamin Marie
LIMSI-CNRS, Orsay, France
Lingua et Machina, Le Chesnay, France
benjamin.marie@limsi.fr

Aurélien Max
LIMSI-CNRS, Orsay, France
Univ. Paris Sud, Orsay, France
aurelien.max@limsi.fr

Abstract

We introduce *pre-post-editing*, possibly the most basic form of interactive translation, as a touch-based interaction with iteratively improved translation hypotheses prior to classical post-editing. We report simulated experiments that yield very large improvements on classical evaluation metrics (up to 21 BLEU) as well as on a parameterized variant of the TER metric that takes into account the cost of matching/touching tokens, confirming the promising prospects of the novel translation scenarios offered by our approach.

1 Introduction

As shown by oracle studies (Wisniewski et al., 2010; Turchi et al., 2012; Marie and Max, 2013), Statistical Machine Translation (SMT) systems produce results that are of significantly lower quality than what could be produced from their available resources. As a pragmatic solution, human intervention is commonly used for improving automatic draft translations, in so-called *post-editing* (PE), but is also studied earlier in the translation process in a variety of interactive strategies, including e.g. completion assistance and local translation choices (e.g. (Foster et al., 2002; Koehn and Haddow, 2009; González-Rubio et al., 2013)). Although *interactive machine translation* does facilitate the work of the SMT system in certain situations by allowing it to make efficient use of knowledge contributed by the human translator, post-editing has been shown to remain a faster alternative (Green et al., 2014). Nevertheless, this activity usually requires complex intervention from an expert translator (Carl et al., 2011).

In this work we reduce interaction with an SMT system to its most basic form: similarly to what a human translator is likely to do when first reading a draft translation to post-edit, we require a user to simply *spot* those segments of a draft translation that can participate in an acceptable translation. The corresponding information is then used by a SMT system in a soft way to improve the draft translation. This process may be iteratively repeated as long as enough improvements are obtained, and terminates with classical post-editing on the obtained translation, hence we dub it *pre-post-editing* (PPE). We resort to simulated pre-post-editing and post-editing, as in other works (Carl et al., 2011; Denkowski et al., 2014), to measure translation performance on some available reference translation using both classical metrics and a variant of the TER metric (Snover et al., 2006), where, essentially, the cost of a token *matching* operation is a parameterized fraction of the cost of the other token edit operations. With the implementation of appropriate strategies in the SMT system, we show under reasonable assumptions that this approach has the potential to significantly reduce the amount of human effort required to obtain a final translation.

In the remainder of this article, we describe the technical details of pre-post-editing (Section 2), report experiments conducted on two translation directions and two domains (Section 3), and finally discuss our proposal and introduce our future work (Section 4).

2 Touch-based pre-post-editing

In our PPE framework, the human *pre-post-editor* has to mark n-grams from a translation hypothesis that can take part in a correct translation.[1] The annotated n-grams are counted, as an n-gram can appear more than once in the same sentence, and a "positive" 6-gram language model (LM)

[1] A touch-based interface when a keyboard is not available or typing is inconvenient lends itself particularly well to PPE.

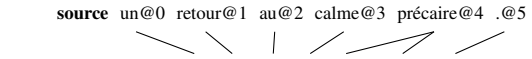
Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1040–1045,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

(positive-lm) is trained on these counts[2]. A "negative" LM (negative-lm) is also trained on the counted n-grams left unannotated. Then, all bi-phrases from the SMT system's phrase table that match an annotated n-gram, according to the source token alignments provided by the decoder, are removed from the main phrase table and stored in a separate "positive" phrase table (positive-pt). Conversely, n-grams containing at least one token left unannotated are considered as incorrect, and the set of bi-phrases matching these n-grams are removed and stored in a "negative" phrase table (negative-pt).

As source tokens can appear more than once in a source text, they are *located*: an identifier is concatenated to each token to make it unique in the source text. Tokens of the source phrases in the phrase table are accordingly also located, so each bi-phrase is duplicated as needed to cover all located tokens. Using located tokens allows our PPE framework to treat differently source tokens that are correctly translated from incorrectly translated ones in the same sentence or text. Figure 1 shows an example of phrase table extraction, using located source tokens[3], for one iteration of PPE.

If an n-gram is annotated as correct, all its inner n-grams of lower order are also deemed correct. Although annotating translations of high quality may be less expensive by explicitely annotating *incorrect* n-grams instead of correct ones, such annotations would not permit to identify correct n-grams inside incorrect ones, as illustrated in Figure 2. PPE can thus be worded as a simple problem for the pre-post-editor: *which sequences of tokens should appear in the final translation?*

The newly extracted phrase tables and LMs[4], along with the remainder of the original phrase table and the original LM, are used to re-decode the source text in a first iteration of PPE. A new PPE annotation can then be performed on the new translations. The newly extracted "positive" and "negative" phrase tables are merged with the corresponding phrase table of the previous iteration. The extracted n-gram counts from the current iteration and the counts of the previous iterations are summed, and the LMs are re-trained with the updated counts. A new iteration of PPE is then performed with the updated models. The weights for all, old or new, models in the log-linear combination are found by tuning on a development set for each PPE iteration.[5] Figure 3 illustrates 4 iterations of PPE from an initial translation hypothesis assuming a given target reference translation.



positive-pt source	positive-pt target	negative-pt source	negative-pt target
retour@1 au@2 précaire@4 .@5	return to precarious .	précaire@4 calme@3 précaire@4 précaire@4 .@5	is precarious calm is precarious is precarious .

positive-lm n-gram	positive-lm count	negative-lm n-gram	negative-lm count
return	1	a	1
return to	1	a return to	1
to	1	to precarious	1
calm	1	to calm is precarious .	1

Figure 1: Examples of some of the bi-phrases and n-grams extracted for phrase tables and language models according to a reference translation.

source	son impopularité semble être en grande partie due au chômage
PPE#0	his unpopularity seems to be / owing / largely / to unemployment
PPE#1	his unpopularity seems to be largely owing to unemployment
target ref	his unpopularity seems to be largely owing to unemployment

Figure 2: Annotation example for two correct tokens forming an incorrect n-gram. At the first PPE iteration a reordering is performed and the new hypothesis now matches the reference translation.

3 Experiments

3.1 Data and systems

We ran experiments on two translation tasks of different domains: the WMT'14 Medical translation task (medical) and the WMT'11 news translation task (news) for the language pair en-fr on both translation directions. For both tasks we trained two competitive phrase-based SMT systems using Moses (Koehn et al., 2007) and WMT data[6] (see Table 1). The tuning for all systems, including our iteration-specific PPE systems, was performed with kb-mira (Cherry and Foster, 2012).

3.2 An adapted evaluation metric: TER_{PPE}

Classical MT evaluation metrics cannot take into account the interactive cost of PPE, and thus do

[2] We used SRILM (Stolcke, 2002) to train the LMs with Witten-Bell smoothing.

[3] Subsequent examples do not use located tokens.

[4] The extracted LMs are sentence-level, and are only used on their specific sentence during PPE.

[5] In this work, we did not exceed 5 iterations.

[6] http://www.statmt.org/wmt14

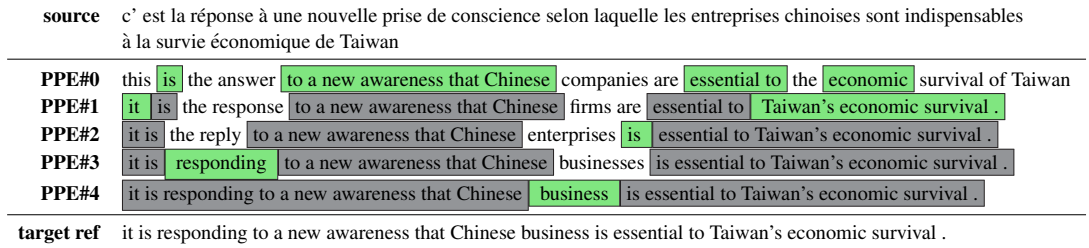


Figure 3: Example of a pre-post-edition trace for French to English translation (using the `news` task, cf. Section 3) using a given implicit target reference translation for simulating pre-post-editing and post-editing. Each newly touched phrase is indicated with a green background. Phrases with a gray background indicate previously touched phrases but their tokens remain individually touchable by the user.

Tasks	Corpus	Sentences	Tokens (fr-en)
`news`	train	12M	383M - 318M
	dev	2,525	73k - 65k
	test	3,003	85k - 74k
`medical`	train	4.9M	91M - 78M
	dev	500	12k - 10k
	test	1,000	26k - 21k
	specialized LM		146M - 78M
for both tasks	LM		2.5B - 6B

Table 1: Data used in this work.

not allow us to make direct comparisons with PE. We thus adapt the TER (Snover et al., 2006) metric, which typically uses 4 types of token edits: `substitution` (s), `insertion` (i), `deletion` (d) and `shift` (f). While these edit types all have a (debatable) uniform cost of 1, the operation of `matching` (m) a correct token is ignored. We posit that this operation is in fact performed by a human translator during PE (at the minimum, by recognizing and skipping tokens), and that it can be directly compared to our touch-based selection of tokens for PPE. However, we cannot at this stage of our work provide a realistic cost value for this operation, and so we introduce a match cost parameter α, and use the following as our PPE-aware metric:

$$\text{TER}_{\text{PPE}} = \frac{\#s + \#d + \#i + \#f + \alpha\#m}{r + \alpha r} \quad (1)$$

where r is the number of tokens in the reference translation. Note that a null value for α makes TER_{PPE} correspond to TER, while a value of 1 would indicate that a token matching/touch (m) is e.g. as costly as a token rewriting (s). We anticipate that a realistic value for α given a reasonably skilled user should be rather small, but we will provide TER_{PPE} results for the full range $[0, 1]$.

3.3 Experimental results

To validate our approach, we initially used a simulated post-editing paradigm (Carl et al., 2011; Denkowski et al., 2014) in which non-post-edited reference translations are used in lieu of human post-editions. Results on TER (Snover et al., 2006) and BLEU (Papineni et al., 2002), tuning on both metrics, are provided in Tables 2 (`news`) and 3 (`medical`).

First, we observe that whatever the metric and the task, the first iteration of PPE always yields a significant improvement over the `Moses` initial system (e.g. up to +9.8 BLEU and -8.2 TER for `news` fr→en). Unsurprisingly, tuning on a metric yields better results for the same metric for the first iteration; however, we note that this is not always true for the TER metric at later iterations (cf. `news` en→fr). More generally, tuning on the TER metric results in lower improvements for `news`, which are mostly concentrated on the first iterations; as systems tuned on BLEU have been found to produce better translations than systems tuned on TER (Cer et al., 2010), only BLEU tuning was used for `medical`.[7]

Improvements follow an interesting pattern over PPE iterations: for instance, on `news` fr→en, BLEU scores steadily increase after each new touch-based iteration and reach a gain of +21.1 BLEU and -12.3 TER over the initial `Moses` translation after 5 PPE iterations. Results are very comparable on both language pairs and both domains, e.g. gains of +12.1 BLEU and -9.7 TER are obtained on fr→en `medical`. The lesser amplitude of the gains obtained after 5 iterations may be attributed to the higher ini-

[7] We have observed a tendency of the TER tuning to shrink the size of hypotheses, resulting in higher brevity penalty values for BLEU and a higher number of insertions for TER.

Iteration	fr→en				en→fr			
	tuned with TER		tuned with BLEU		tuned with TER		tuned with BLEU	
	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU
Moses	51.1	28.2	52.7	28.6	52.3	29.7	51.8	31.1
PPE iteration 1	42.9	35.4	46.7	38.4	44.4	35.0	47.3	39.6
PPE iteration 2	40.8	37.3	43.7	43.4	43.0	36.3	44.6	43.9
PPE iteration 3	40.8	37.8	42.2	46.2	42.5	36.4	43.5	46.6
PPE iteration 4	39.9	37.9	40.9	48.3	42.3	36.5	42.3	48.2
PPE iteration 5	39.9	37.9	40.4	49.7	42.2	36.6	41.0	49.5

Table 2: PPE results on the news task.

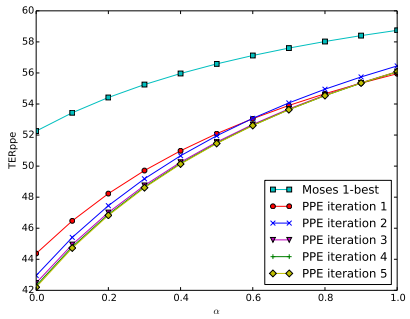
tial quality of the translations in the medical task (e.g. 37.1 BLEU vs 28.6 BLEU in fr→en for Moses with BLEU tuning).

Iteration	fr→en		en→fr	
	tuned with BLEU		tuned with BLEU	
	TER	BLEU	TER	BLEU
Moses	42.2	37.1	44.0	38.8
PPE iteration 1	36.9	44.9	37.2	48.3
PPE iteration 2	34.8	47.5	35.3	51.1
PPE iteration 3	34.1	48.5	33.5	52.9
PPE iteration 4	32.9	49.2	32.4	54.0
PPE iteration 5	32.5	49.2	32.1	54.8

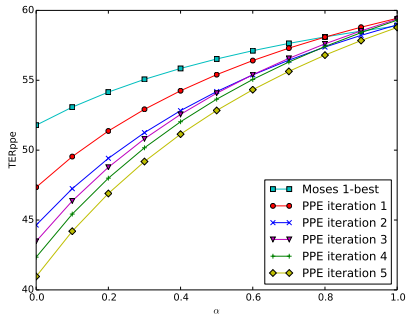
Table 3: PPE results on the medical task.

Figures 4 and 5 show how our TER_{PPE} metric varies for different values of our α parameter (recall that $\alpha = 0$ corresponds to TER). Essentially, whatever the value of α, we observe that any iteration of PPE dominates PE (Moses 1-best), but with a tendency to become as costly as PE for high, but probably unrealistic values of α. Tuning with BLEU allows us to bring regular improvements as the number of iteration increases, while tuning with TER makes the amplitude of the gains decrease faster.

Furthermore, results shown in Table 4 point out the complementarity between negative models (negative-lm and negative-pt) and positive models (positive-lm and positive-pt), with a drop of almost 10 BLEU points compared to the corresponding configuration using all models when removing one type of models on both translation directions. The language models (negative-lm and positive-lm) seem to play a more important role during PPE than the phrase tables (negative-pt and positive-pt), with a drop of 9.6 BLEU points on news fr→en when removing the language models against a significantly lower drop of 4.4 BLEU points when removing the phrase tables.



(a) Tuned with TER



(b) Tuned with BLEU

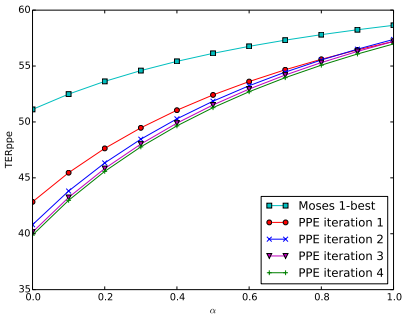
Figure 4: PPE results on the en→fr news task.

4 Discussion and future work

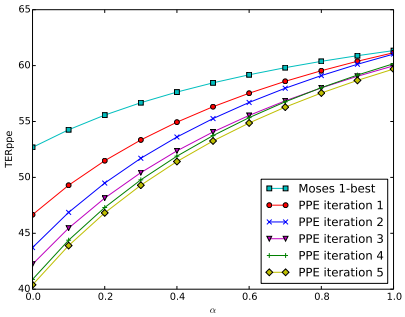
We have introduced *pre-post-editing*, a minimalist interactive machine translation paradigm where a user is only asked to spot text fragments that may be used in the final translation. Our approach is quite comparable to the two-pass procedure described by Luong et al. (2014) using word-level confidence estimation (e.g. (Bach et al., 2011)) to update the cost of the search graph hypotheses. However, contrarily to Luong et al.'s work, our PPE framework is efficiently multi-pass, updates the models over iterations and relies on more informative annotations made at n-gram-level. Our evaluation based on simulated post-editing has revealed a large potential for translation improvement. Interestingly, the type of interaction defined

Configuration	fr→en tuned with BLEU TER	fr→en tuned with BLEU BLEU	en→fr tuned with BLEU TER	en→fr tuned with BLEU BLEU
`Moses`	52.7	28.6	51.8	31.1
PPE w/ all models	40.4	49.7	41.0	49.5
PPE w/o `negative-pt` and `negative-lm`	45.2	39.4	47.1	39.0
PPE w/o `positive-pt` and `positive-lm`	46.7	39.8	48.3	39.8
PPE w/o `negative-pt` and `positive-pt`	45.0	45.3	46.5	44.9
PPE w/o `negative-lm` and `positive-lm`	42.7	40.1	43.2	42.0

Table 4: PPE results for the `news` task after 5 iterations using various configurations.



(a) Tuned with TER



(b) Tuned with BLEU

Figure 5: PPE results on the fr→en `news` task.

is very different from that expected of a post-editor or in existing interactive translation modes, and lends itself nicely to touch-based interaction. Furthermore, our proposal may in fact define a new *role* in Computer-Assisted Translation, with PPE being performed on-the-go on mobile devices by more people than available human translators, and even possibly by monolinguals of the target language whose contribution may be more efficiently exploited than that of monolinguals of the source language (e.g. (Resnik et al., 2010)).

In terms of usability, our future work will focus on two important questions: (*a*) study the actual use of PPE in an interactive setting and tune the α parameter for our TER_{PPE} metric on HTER (Snover et al., 2006) traces, and (*b*) study whether PPE alters in any positive way the work of the human translator performing the residual post-editing, hoping that PE *could become a less tedious task by nature*. We further anticipate that some additions would improve our approach, including dealing early with out-of-vocabulary phrases, proposing local drop-down options (e.g. (Koehn and Haddow, 2009)), possibly clustered by senses, allowing the user to easily fix reordering issues, and adapting PPE to be discourse-aware (e.g. (Ture et al., 2012)).

5 Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. The work of the first author is supported by a CIFRE grant from French ANRT.

References

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: A Method for Measuring Machine Translation Confidence. In *Proceedings of ACL*, Portland, USA.

Michael Carl, Dragsted Barbara, Jakob Elming, Hardt Daniel, and Jakobsen Arnt Lykke. 2011. The Process of Post-Editing: A pilot study. In *Copenhagen Studies in Language*.

Daniel Cer, Christopher D. Manning, and Daniel Jurafsky. 2010. The best lexical metric for phrase-based statistical mt system optimization. In *Proceedings of NAACL*, Los Angeles, USA.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of NAACL*, Montréal, Canada.

Michael Denkowski, Chris Dyer, and Alon Lavie. 2014. Learning from Post-Editing : Online Model Adaptation for Statistical Machine Translation. In *Proceedings of EACL*, Gothenburg, Sweden.

George Foster, Philippe Langlais, and Guy Lapalme. 2002. User-Friendly Text Prediction for Translators. In *Proceedings of EMNLP*, Philadelphia, USA.

Jesús González-Rubio, Daniel Ortiz-Martínez, José-Miguel Benedi, and Francisco Casacuberta. 2013. Interactive Machine Translation using Hierarchical Translation Models. In *Proceedings of EMNLP*, Seattle, USA.

Spence Green, Sida Wang, Jason Chuang, Jeffrey Heer, Sebastian Schuster, and Christopher D. Manning. 2014. Human Effort and Machine Learnability in Computer Aided Translation. In *Proceedings of EMNLP*, Doha, Qatar.

Philipp Koehn and Barry Haddow. 2009. Interactive assistance to human translators using statistical machine translation methods. In *Proceedings of MT Summit*, Ottawa, Canada.

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 ACL, demo session*, Prague, Czech Republic.

Benjamin Marie and Aurélien Max. 2013. A Study in Greedy Oracle Improvement of Translation Hypotheses. In *Proceedings of IWSLT*, Heidelberg, Germany.

Luong Ngoc Quang, Laurent Besacier, and Lecouteux Benjamin. 2014. An Efficient Two-Pass Decoder for SMT Using Word Confidence Estimation. In *Proceedings of EAMT*, Dubrovnik, Croatia.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, Philadelphia, USA.

Philip Resnik, Olivia Buzek, Chang Hu, Yakov Kronrod, Alex Quinn, and Benjamin B. Bederson. 2010. Improving Translation via Targeted Paraphrasing. In *Proceedings of EMNLP*, Cambridge, USA.

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*, Cambridge, USA.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of ICSLP*, Denver, USA.

Marco Turchi, Tijl De Bie, Cyril Goutte, and Nello Cristianini. 2012. Learning to Translate: A Statistical and Computational Analysis. *Advances in Artificial Intelligence*.

Ferhan Ture, Douglas W. Oard, and Philip Resnik. 2012. Encouraging Consistent Translation Choices. In *Proceedings of NAACL*, Montréal, Canada.

Guillaume Wisniewski, Alexandre Allauzen, and François Yvon. 2010. Assessing Phrase-Based Translation Models with Oracle Decoding. In *Proceedings of EMNLP*, Cambridge, USA.