Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language xxx (2014) xxx–xxx

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

# Translating without in-domain corpus: Machine translation post-editing with online learning techniques☆

Antonio L. Lagarda [a,*], Daniel Ortiz-Martínez [b,1], Vicent Alabau [b,1], Francisco Casacuberta [b,1]

[a] *Institut Tecnològic d'Informàtica, Universitat Politècncnica de València, Camí de Vera s/n, 46022 València, Spain*
[b] *PRHLT Research Center, Universitat Politècnica de València, Camí de Vera s/n, 46022 València, Spain*

Received 21 March 2014; received in revised form 19 October 2014; accepted 27 October 2014

**Abstract**

Globalization has dramatically increased the need of translating information from one language to another. Frequently, such translation needs should be satisfied under very tight time constraints. Machine translation (MT) techniques can constitute a solution to this overly complex problem. However, the documents to be translated in real scenarios are often limited to a specific domain, such as a particular type of medical or legal text. This situation seriously hinders the applicability of MT, since it is usually expensive to build a reliable translation system, no matter what technology is used, due to the linguistic resources that are required to build them, such as dictionaries, translation memories or parallel texts. In order to solve this problem, we propose the application of automatic post-editing in an online learning framework. Our proposed technique allows the human expert to translate in a specific domain by using a base translation system designed to work in a general domain whose output is corrected (or adapted to the specific domain) by means of an automatic post-editing module. This automatic post-editing module learns to make its corrections from user feedback in real time by means of online learning techniques. We have validated our system using different translation technologies to implement the base translation system, as well as several texts involving different domains and languages. In most cases, our results show significant improvements in terms of BLEU (up to 16 points) with respect to the baseline systems. The proposed technique works effectively when the n-grams of the document to be translated presents a certain rate of repetition, situation which is common according to the document-internal repetition property.


*Keywords:* Machine translation; Statistical machinetranslation; Interactive machine translation; Automaticpost-editing; Online learning

## 1. Introduction

Globalization has urged the need for high-quality translations with fast turn-around times. Examples of that are companies aiming to internationalize their businesses in order to discover new markets and gain competitive advantage,

☆ This paper has been recommended for acceptance by R.K. Moore.

* Corresponding author. Tel.: +34 96 387 70 69.
*E-mail addresses:* alagarda@iti.upv.es (A.L. Lagarda), dortiz@prhlt.upv.es (D. Ortiz-Martínez), valabau@prhlt.upv.es (V. Alabau), fcn@prhlt.upv.es (F. Casacuberta).
[1] Tel.: +34 96 387 81 70.

http://dx.doi.org/10.1016/j.csl.2014.10.004
0885-2308/

or transnational institutions that have legal requirements to produce documentation in multiple languages. Frequently, these documents need to be delivered with tight deadlines and, at the same time, clients are pushing to adjust prices. As a result, translation agencies and in-house translation departments have been compelled to adopt automated *machine translation* (MT) in an attempt to improve their translation pipelines (Dove et al., 2012). In that way, MT systems are used to produce drafts of the translations that later are post-edited by human translators in order to achieve the high-quality standards required by the industry.

Historically, *rule based machine translation* (RBMT) systems have been used by companies to automate their translation needs (Silva, 2012). Nevertheless, RBMT systems are expensive to personalize, as expert linguists are needed to create bilingual dictionaries or specific rules (Bennett and Slocum, 1985; Isabelle et al., 2007). As a result, these systems are only available for a handful of European languages. On the contrary, *statistical machine translation* (SMT) systems are created in a more unattended manner by harvesting parallel segments from a collection of Bi-texts or *translation memories* (TM). The quality that SMT systems achieve is often better than that of RBMT systems (Béchara et al., 2012; Silva, 2012), at least for some language pairs, and provided that there is enough data. However, it is only recently that SMT systems are being effectively used to improve the productivity of human translators by means of building engines customized from the client's data. Unfortunately, clients seldom have previous parallel corpora from the same domain that can be used to train these customized engines, or to adapt the domain of a pre-existent one (Irvine et al., 2013). Additionally, training such engines may take hours, days, if not weeks of computation. On the other hand, RBMT systems can be used right out-of-the-box, and they can be enhanced with an *automatic post-editing* (APE) by an SMT system in a way that translators appreciate it more than either of both systems alone, regardless their BLEU scores (Béchara et al., 2012). That paper shows that, although automatic evaluation metrics favor the pure SMT system, human evaluators prefer the output provided by the statistically post-edited RBMT system.

Thus, the premise of this work is based on a real case scenario: a human translator, probably a freelancer, is given a translation assignment with a tight turn-around time. Alas, our translator lacks the necessary linguistic resources such as TMs or parallel texts that would allow him or her to build an MT system (no matter which technology is used) adapted to the specific domain of the document. Under these circumstances, what are his or her alternatives?

W/O RESOURCES This is the traditional manual method, but it requires more time and effort. Note that, in this case, we are not considering the use of previously collected TMs neither TMs generated on the go. On the contrary, each sentence is supposed to be translated from an empty box, or filled up with the source text at most.

WEB Translating with a web-based translation application, and then post-edit its output. Nowadays, there are many free web-based translation applications which can achieve a translation quality enough for gisting and, even in some cases, the quality can be satisfactory. However, it can be insufficient for many domains of interest. Also, these web-based translation applications can present some confidentiality issues that should be considered, because all content uploaded will be employed to enrich their models. Moreover, some of them are not free when translating more than a given quantity of words.

RBMT Translating with a RBMT system, and then post-editing its output. There are many RBMT translation systems, some of them free. Nevertheless, the output of RBMT systems is usually not tailored to the domain of the document being translated and fail to adapt to new domains (Isabelle et al., 2007), e.g., lexical choices may not be appropriate. Although APE may alleviate this problem, still parallel corpora is needed.

SMT If he or she is familiar with SMT, he or she can train an SMT model with unrelated corpora (remember that there are no available in-domain TMs, which is a frequent case). As in the RBMT case, these SMT translations will contain several mistakes due to the fact that, in this case, the training corpus is out-of-domain.

Neither of these options is optimal since, as we have discussed above, MT customization is key to improve the translator's productivity. In this paper, we propose a technique to help the translator in this regard. We assume that the translator will adopt one of the different MT alternatives proposed previously as a draft for post-editing, none of which is customized to the document domain. APE can be specially useful under these circumstances since it can be used as a domain adaptation technique. Domain adaptation has received extensive attention from the SMT research community during the last years. However, this topic has typically been approached in scenarios where the set of training samples used to estimate the model parameters (both in and out-of-domain) are available beforehand, and the system does not get updated after the training stage has concluded.



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

In this work the domain adaptation problem is tackled in a different way, specifically, as translation hypotheses are amended by the user, the system will learn from these corrections, using them to train APE models on-the-fly. In this way, the following system translation hypotheses automatically will apply past user's amendments. To achieve this effect, an *on-line learning* (OL) SMT system will be used to customize the output of the original system after each translated segment. Hence, APE is considered here as an online technique, where the user interacts with the system in order to correct the initial hypotheses. In addition to this, these corrections could be taken as new corpora to retrain the APE models. In an ideal APE scenario, these models should be updated for each new hypothesis given by the system, in order to minimize the user post-editing effort. This is why we propose APE as a natural application field of OL techniques.

The rest of the paper is organized as follows. First, Section 2 introduces the techniques implemented in our system, which applies statistical machine translation (Section 2.1) to an automatic post-editing task (Section 2.2) in an online learning environment (Section 2.3). Then, Section 3 compares our proposal with some similar techniques appearing in the literature. Section 4 describes the proposed system, emphasizing the base translation systems (Section 4.1) and our OL approach (Section 4.2). Finally, we show some experiments in Section 5 that apply these techniques, employing corpora with different features (Section 5.1) and several base translation systems (Section 5.2). Finally, we discuss the results in Section 5.3, showing in Appendix A some plots with the dynamics of OL APE.

## 2. Framework

In this section we describe the theoretical foundations of our proposal, where we apply online learning techniques to an automatic post-editing task based on statistical machine translation. Thus, we introduce the statistical approach to machine translation (Section 2.1) as well as two language technologies that can be built upon it: automatic post-editing (Section 2.2), and online learning (Section 2.3).

### 2.1. Statistical machine translation

Given a sentence f from a source language $\mathcal{F}$ to be translated into a target sentence e of a target language $\mathcal{E}$, the fundamental equation of SMT (Brown et al., 1993) is the following:

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{argmax}\,\{Pr(\mathbf{e} \mid \mathbf{f})\} \tag{1}$$

$$= \underset{\mathbf{e}}{argmax}\,\{Pr(\mathrm{f} \mid \mathbf{e})\, Pr(\mathbf{e})\} \tag{2}$$

where $Pr(\mathbf{f} \mid \mathbf{e})$ is approximated by a *translation model* that represents the correlation between the source and the target sentence and where $Pr(\mathbf{e})$ is approximated by a *language model* representing the well-formedness of the candidate translation e.

State-of-the-art statistical machine translation systems follow a log-linear approach (Och and Ney, 2002), where direct modelling of the posterior probability $Pr(\mathbf{e} \mid \mathbf{f})$ of Eq. (1) is used. In this case, the decision rule is given by the expression:

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{argmax} \left\{ \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{f}) \right\} \tag{3}$$

where each $h_m(\mathbf{e}, \mathbf{f})$ is a feature function representing a statistical model and $\lambda_m$ its weight.

Current most popular MT systems are based on the use of phrase-based models (Koehn et al., 2003) as translation models. The basic idea of phrase-based translation is to segment the source sentence into phrases, then to translate each source phrase into a target phrase, and finally to reorder the translated target phrases in order to compose the target sentence. If we summarize all the decisions made during the phrase-based translation process by means of the hidden variable $\tilde{a}_1^K$, we obtain the expression:

$$Pr(\mathbf{f} \mid \mathbf{e}) = \sum_{K, \tilde{a}_1^K} Pr(\tilde{f}_1^K, \tilde{a}_1^K \mid \tilde{e}_1^K) \tag{4}$$



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

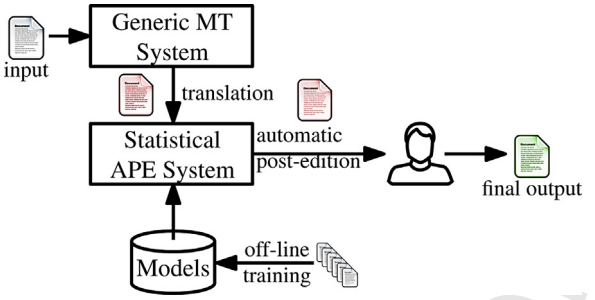



Fig. 1. Diagram of an APE system.

where each $\tilde{a}_k \in \{1 \ldots K\}$ denotes the index of the target phrase $\tilde{e}$ that is aligned with the $k$-th source phrase $\tilde{f}_k$, assuming a segmentation of length $K$.

According to Eq. (4), and following a maximum approximation, the problem stated in Eq. (2) can be re-framed as:

$$\hat{\mathbf{e}} \approx \underset{\mathbf{e},\mathbf{a}}{arg\max}\{p(\mathbf{e}) \cdot p(\mathbf{f}, \mathbf{a} \mid \mathbf{e})\} \tag{5}$$

Following the log-linear approach stated in Eq. (3), Eq. (5) can be rewritten as:

$$\hat{\mathbf{e}} = \underset{\mathbf{e},\mathbf{a}}{argmax} \left\{ \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{a}, \mathbf{f}) \right\} \tag{6}$$

which is the approach that we follow in this work.

### 2.2. Automatic post-editing

MT systems usually need a final revision step by a human post-editor, to assure a quality output. This can be a tedious task, where the post-editor will have to repeatedly correct the same mistakes, due to the systematic behavior of MT systems (Allen and Hogan, 2000; Carpuat and Simard, 2012).

This correction process can be understood as a transformation from an input (the translation provided by the previous MT system, usually with errors), to an output (a text in the same language where those errors have been amended). Thus, post-editing could be considered as a translation between two languages. APE systems were proposed by Knight and Chander (1994) to try to automate as far as possible that final human revision phase. Some authors consider APE as a domain adaptation or customization technique (Isabelle et al., 2007; Diaz et al., 2008; Rubino et al., 2012).

In a statistical APE system, SMT models are trained to correct the outputs of another MT system (Simard et al., 2007), which is often considered as a black box. In this way, the fundamental equation of SMT (Eq. (1)) would be applied from a sentence $\mathbf{e}'$ of target language with errors $\mathcal{E}'$, which is the output of the previous MT system that need to be corrected, into a target sentence e of a target language $\mathcal{E}$ without errors (hopefully, at least with fewer errors than $\mathcal{E}'$).

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{argmax} \left\{ Pr(\mathbf{e}|\mathbf{e}') \right\} \tag{7}$$

Fig. 1 shows a diagram of a statistical APE system. In the diagram, the source sentences are processed as input by a generic MT system, producing a set of translations. After that, each system translation is used to feed the statistical APE system, whose models have been initialized from parallel corpus that have also been translated with the same MT system. For a given system translation, the APE system produces an automatic post-edition that is corrected by the user to generate the final output.

### 2.3. Online learning for SMT

One key feature of the technique proposed in this paper is the application of online learning. Here we describe the concept of online learning, including a brief review of different works that apply this concept to SMT (Section 2.3.1),

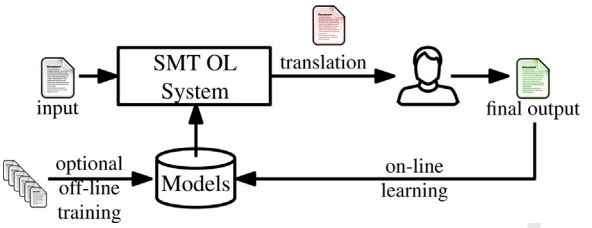



Fig. 2. Diagram of an SMT system with OL.

the specific formulation of the log-linear SMT model with online learning used in this paper (Section 2.3.2) as well as a brief description of how the models are extended from new training samples (Section 2.3.3).

### 2.3.1. Concept of online learning

Online learning is a machine learning task that is structured in a series of trials, where each trial has three steps: (1) the learning algorithm receives an instance, (2) a label for the instance is predicted and (3) the true label for the instance is presented.

Online learning fits nicely in MT post-editing tasks, since the output of MT systems is corrected and validated by expert translators. If an SMT system is used to generate the translations, then its statistical models can be modified from the newly generated training pairs. Fig. 2 shows a diagram of an SMT system with OL.

During the last years, there has been an increasing interest in developing techniques to adapt or train the features of a log-linear combination in online learning settings. As far as we know, the SMT system with OL proposed by Ortiz-Martínez et al. (2010) (the system used in this paper as it is explained in the following section) constitutes the first work that successfully applies OL to SMT, solving the technical limitations encountered in previous works without the need of introducing heuristic approximations. Such previous works on online SMT include the dynamic adaptation of an IMT system via cache-based model extensions proposed by Nepveu et al. (2004) and the statistical computer assisted translation scenario with online learning proposed by Cesa-Bianchi et al. (2008). In both cases, the proposed systems were heavily limited by their inability to extend the translation models due to technical limitations to efficiently incorporate new parameters in a principled way. The work presented by Hardt and Elming (2010) applies a cache-based strategy similar to that presented by Nepveu et al. (2004), where the translation model is extended by means of heuristic IBM4-based word alignment techniques. IBM-4 word alignment techniques are replaced by phrase alignment techniques to extend the translation model in Bertoldi et al. (2013), Wäeschle et al. (2013). An additional attempt to efficiently extend the translation model was proposed by Blain et al. (2012). Their proposal aligns the output of the decoder with the reference given by the user as a previous step to obtain the word alignments between the source and reference sentences that are necessary to extract new phrase pairs. The method used to align the system translation and the reference sentence is based on the edit distance algorithm since it is assumed that both sentences will be similar.

### 2.3.2. A log-linear SMT model with online learning

Here we adopt the online learning techniques described by Ortiz-Martínez et al. (2010). In that work, the authors define an incrementally updateable SMT model for its application in the interactive machine translation framework. Such an SMT model is able to process new training samples one by one, with constant computational complexity (i.e. the complexity does not depend on the training samples that have been previously seen). Moreover, their proposed system has already been implemented in a certain number of SMT prototypes (Ortiz-Martínez et al., 2011; Alabau et al., 2014).

The SMT system described by Ortiz-Martínez et al. (2010) uses a log-linear model to generate its translations. According to Eq. (6), we introduce a set of seven feature functions (from $h_1$ to $h_7$): a $n$-gram language model ($h_1$), an inverse sentence-length model ($h_2$), inverse and direct phrase-based models ($h_3$ and $h_4$ respectively), a target phrase-length model ($h_5$), a source phrase-length model ($h_6$), and a distortion model ($h_7$). The details for each feature function are listed below:



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

- **$n$-gram language model** ($h_1$):
  $h_1(\mathbf{e}) = \log(\prod_{i=1}^{|\mathbf{e}|+1} p(e_i \mid e_{i-n+1}^{i-1}))$,[2] $h_1$ can be implemented by means of smoothed $n$-gram language models. Here we adopt an interpolated $n$-gram model with Kneser-Ney smoothing.
- **source sentence-length model** ($h_2$): $h_2(\mathbf{f}, \mathbf{e}) = \log(p(|\mathbf{f}| \mid |\mathbf{e}|))$, $h_2$ can be implemented by means of a set of Gaussian distributions whose parameters are estimated for each source sentence length.
- **inverse and direct phrase-based models** ($h_3$, $h_4$): $h_3(\mathbf{e}, \mathbf{a}, \mathbf{f}) = \log(\prod_{k=1}^{K} p(\tilde{f}_k \mid \tilde{e}_{\tilde{a}_k}))$, where $h_3$ is implemented with an inverse phrase-based model. This phrase-based model is smoothed with an HMM-based alignment (Vogel et al., 1996) model by means of linear interpolation.
  Analogously $h_4$ is defined as: $h_4(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \log(\prod_{k=1}^{K} p(\tilde{e}_{\tilde{a}_k} \mid \tilde{f}_k))$
- **target phrase-length model** ($h_5$): $h_5(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \log(\prod_{k=1}^{K} p(|\tilde{e}_k|)$, this feature is modelled by means of a geometric distribution. The geometric distribution penalizes the length of the target phrases.
- **source phrase-length model** ($h_6$): $h_6(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \log(\prod_{k=1}^{K} p(|\tilde{f}_k| \mid |\tilde{e}_{\tilde{a}_k}|))$, a geometric distribution can be used to model $h_6$, such distribution penalizes the difference between the source and target phrase lengths.
- **distortion model** ($h_7$): $h_7(\mathbf{a}) = \log(\prod_{k=1}^{K} p(\tilde{a}_k \mid \tilde{a}_{k-1}))$, again, this feature function can be modelled by means of a geometric distribution. Such distribution penalizes the re-orderings.

### 2.3.3. *Extending the log-linear model from user feedback*

In order to incrementally train the log-linear model, a set of sufficient statistics that can be incrementally updated should be maintained for each feature function. If the estimation of the statistical model does not require the use of the expectation–maximization (EM) algorithm (Dempster et al., 1977) (e.g. n-gram language models), then it is generally easy to incrementally update the model given a new training sample. By contrast, if the EM algorithm is required (e.g. word alignment models), the estimation procedure has to be modified, since the conventional EM algorithm is designed for its use in batch learning scenarios. For those models, the incremental version of the EM algorithm (Neal and Hinton, 1999) is applied. Incremental EM guarantees estimation convergence after each algorithm iteration in a similar way to conventional EM, but E and M steps are individually applied to each training sample.

In this section we identify the sufficient statistics for the main components used in the log-linear combination described above. These components are the language model (feature $h_1$) and the translation model (features $h_3$ and $h_4$). Source and target phrase-length models (features $h_5$ and $h_6$) and the distortion model (feature $h_7$) are implemented by means of geometric distributions with fixed parameters and thus they do not require a complex treatment. Finally, since the sentence length model (feature $h_2$) is implemented by means of gaussian distributions, well known incremental update rules using simple sufficient statistics can be found in the literature (see for instance Knuth (1981)).

*Sufficient statistics for the language model ($h_1$)* Since language models are implemented using interpolated Kneser-Ney smoothing, probabilities are generated according to the following equation:

$$p(e_i \mid e_{i-n+1}^{i-1}) = \frac{\max\{c_X(e_{i-n+1}^{i}) - D_n, 0\}}{c_X(e_{i-n+1}^{i-1})} + \frac{D_n}{c_X(e_{i-n+1}^{i-1})} N_{1+}(e_{i-n+1}^{i-1}\bullet) \cdot p(e_i \mid e_{i-n+2}^{i-1}) \tag{8}$$

where $D_n = (c_{n,1})/(c_{n,1} + 2c_{n,2})$ is a fixed discount ($c_{n,1}$ and $c_{n,2}$ are the number of $n$-grams with one and two counts respectively), $N_{1+}(e_{i-n+1}^{i-1}\bullet)$ is the number of unique words that follows the history $e_{i-n+1}^{i-1}$ and $c_X(e_{i-n+1}^{i})$ is the count of the $n$-gram $e_{i-n+1}^{i}$, where $c_X(\cdot)$ can represent true counts $c_T(\cdot)$ or modified counts $c_M(\cdot)$ (see Chen and Goodman (1996) for more details).

Under these circumstances, the list of incrementally updateable sufficient statistics for the language model will include $c_{k,1}$, $c_{k,2}$, $N_{1+}(\cdot)$, $c_T(\cdot)$ and $c_M(\cdot)$. In this particular case, the EM algorithm is not required, greatly simplifying the update process for a new training sample (see Ortiz-Martínez et al. (2010) for more details).

[2] $|\mathbf{e}|$ is the length of e, $e_0$ denotes the *begin-of-sentence* symbol, $e_{|e|+1}$ is the *end-of-sentence* symbol and $e_i^j \equiv e_i \ldots e_j$.

*Sufficient statistics for the translation model ($h_3$ and $h_4$)* Features $h_3$ and $h_4$ are implemented by means of inverse and direct phrase models. Since phrase-based models are symmetric models, only an inverse phrase-based model is maintained. Inverse phrase model probabilities are obtained from the relative frequencies of a set of phrase pairs:

$$p(\tilde{f} \mid \tilde{e}) = \frac{c(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}'} c(\tilde{f}', \tilde{e})} \tag{9}$$

According to Eq. (9), the set of sufficient statistics for the phrase models is composed of a set of phrase counts $c(\tilde{f}, \tilde{e})$, which following standard estimation techniques, can be extracted from word alignment matrices. More specifically, given a sentence pair and its corresponding word alignment matrix, only those phrase pairs that are *consistent* with such word alignment matrix are extracted (see Koehn et al. (2003) for more details). Because of this, we also need to maintain direct and inverse HMM-based alignment models. These models are not only useful for smoothing purposes (as it was explained in Section 2.3.2), but also for generating the word alignment matrices that are required to obtain the phrase counts. Since the estimation of HMM-based alignment models requires the use of the EM algorithm, here we need to replace conventional EM by its incremental counterpart, allowing us to modify the parameters of the models for each individual training pair (a more detailed description of the incremental estimation of HMM-based alignment models can be found in Ortiz-Martínez et al. (2010)).

## 3. Related work

In the last few years, there have been different proposals involving APE systems. The majority of such proposals employ SMT to automatically post-edit the translations proposed by rule-based systems (Simard et al., 2007; Dugast et al., 2007; Terumasa, 2007; Lagarda et al., 2009; Béchara et al., 2012). However, statistical APE of an SMT system has also proved its effectiveness (Béchara et al., 2011; Oflazer and El-Kahlout., 2007).

On the other hand, some authors consider APE as a domain adaptation or customization technique (Isabelle et al., 2007; Diaz et al., 2008; Rubino et al., 2012). Domain adaptation constitutes an important topic within the SMT field, with many works tackling the problem under different points of view, such as model interpolation or parameter weighting (Koehn and Schroeder, 2007; Foster et al., 2010), data harvesting (Zhao et al., 2004), or data selection (Lü et al., 2007), just to name a few.

This work also embraces the domain adaptation perspective of APE mentioned above, but putting special emphasis on its applicability to online learning settings. In many real translation scenarios, the set of training samples used to estimate the model parameters is not known a priori, instead, there is a stream of training data that grows continually over time (for instance, the documents to be translated at a translation agency). Under these circumstances, the models used by the system can be continually updated so as to improve the quality of the output. This situation clearly differs from conventional domain adaptation scenarios, where there are closed training sets that are used to estimate the parameters of static models.

The most similar work to our own is presented by Simard and Foster (2013), where the authors propose an APE system in which the automatic post-editing module is implemented as an SMT system with OL. This technique learns post-editor corrections and applies them on-the-fly to further MT output. The authors of that paper prove that this method is effective when translating documents with high levels of internal repetition. This situation is not uncommon, since according to Church and Gale (1995), if a segment is to be repeated, this has the greatest chance of happening within the same document where the segment initially appeared. Under these circumstances, if an error occurs repeatedly, the system will be able to learn the correction provided by the user and apply it when necessary in the following sentences to post-edit. The automatic post-editing module is implemented as a phrase translation system trained from the system translations and the final translations validated by the user. To extract the phrase pairs, the proposed system generates word level alignments based on edit distance. The use of word alignments based on edit distance relies on the heuristic assumption that the system translation and the final translation given by the user are similar. However, this assumption may be difficult to justify with complex or real translation tasks, where the quality of the MT system translations may be low, decreasing the similarity between them and the final translations given by the user (one example of this situation occurs when the MT system is implemented using statistical methods and the sentences to be translated contain poorly represented or unseen events).



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

Our work differs from the work described by Simard and Foster (2013) in that it replaces the automatic post-editing module based on edit distance word alignments by a fully functional online SMT system that is able to learn from scratch or from previously estimated models in real time (Ortiz-Martínez et al., 2010). In this online SMT system, word level alignment generation is no longer based on edit distance but on an HMM-based alignment model estimated by means of the incremental version of the EM algorithm (see Section 2.3 for more details), removing the heuristic approximations that are required in Simard and Foster (2013) to update the phrase-based model.

## 4. System description

In our system, first we *machine translate* the source document, and later we post-edit its translation in an OL framework. Let us remember that in the real case scenario proposed in the introduction, the user did not have translation memories, nor similar corpora to translate the text. To simulate this situation, we have chosen some RBMT systems, and some free web-based translation applications. In addition, we have contemplated the case where the user has some unrelated corpora to train an SMT system.

### 4.1. Base translation systems

In this section, we introduce the base translation systems that we have chosen to post-edit in our experiments.

W/O RESOURCES Under the assumption that no translation memories are used, the user can only manually translate or, at most, copy the source text as a starting point. The latter may be useful when untranslatable proper nouns are present or when source and target languages are very similar (e.g., from the same family of languages).

RBMT We have employed two RBMT systems:

- *Apertium* (Forcada et al., 2011), a free/open-source RBMT tool, initially aimed at related-language pairs but recently expanded to deal with more divergent language pairs.
- *PAHOMTS* (Vasconcellos and León., 1985), a proprietary RBMT system developed by the *Pan-American Health Organization* (PAHO), specialized in the translation of life science texts between English, Portuguese and Spanish.

WEB In the last few years, many web-based translation applications have appeared. In spite of the confidential issues discussed in the introduction, they are broadly used by both professional translators and plain users. They combine many translation paradigms, including RBMT, SMT, TM, semantics, etc., covering a wide range of languages. We have selected the following web-based translation applications:

- *Google Translate*, https://translate.google.com/
- *Bing Translator*, http://www.bing.com/translator
- *Yandex Translate*, http://translate.yandex.com/

SMT SMT needs a previous training step where statistical models are learnt from the given corpora. Although in our premise, there are not similar TM nor corpora to train those models, we have taken into account the option of downloading publicly available corpora to train statistical models, in spite of not knowing if they cover the same domains. We have performed these SMT experiments by means of:

- *Moses* (Koehn et al., 2007), an open-source toolkit which implements state-of-the-art SMT techniques, in this case trained with out-of-domain data.

Finally, as an **ORACLE**, we have considered an in-domain SMT case, where *Moses* was trained with in-domain corpora. These experiments are only taken as a reference, as a best-case scenario where the user has found some completely related corpora to train the statistical models. As we have explained in previous sections, this case is not always realistic, because these corpora are frequently not available. However, we wanted to take them into account to know an empirical upper bound for the proposed techniques.



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

Table 1
Main figures of the tokenized version of the corpora, k and M stand for thousands and millions of elements respectively. Perplexity and OOV rates are computed for 5-grams with IRSTLM (Federico et al., 2008). Repetition rate of those n-grams appearing in the test set, and BLEU between source and target sentences in the test set set will be useful to analyze OL performance.

| Corpus | Domain | Partition | Sentences | Tokens | Vocabulary | OOV rate | Perplexity (5-grams) | Repetition rate |
|---|---|---|---|---|---|---|---|---|
| EMEA | Health | Training | 337 k | 5.7 M/5.1 M | 79 k/67 k | –/– | –/– | –/– |
| (es/en) | | Test | 5 k | 84 k/77 k | 8 k/7 k | 0.02/0.022 | 61.4/68.6 | 30.6/31 |
| Xerox | Technical | Training | 56 k | 747 k/662 k | 17 k/15 k | –/– | –/– | –/– |
| (es/en) | | Test | 1 k | 10 k/8 k | 2 k/2 k | 0.022/0.025 | 87/175 | 24.4/19.8 |
| i3media | News | Training | 1.5 M | 35 M/36.2 M | 360 k/366 k | –/– | –/– | –/– |
| (es/ca) | | Test | 5 k | 122 k/126 k | 18.5 k/18 k | 0.006/0.006 | 164.7/140.4 | 11.7/13.2 |
| Europarl-v7 (es/en) | Proceedings | Training | 2 M | 58 M/55 M | 192 k/130 k | –/– | –/– | –/– |

### 4.2. Online learning automatic post-editing system

Once we have translated the source document with one of the previously introduced systems, we will need a post-editing step. In this task, the user will correct the hypotheses given by the MT system in order to achieve the desired final translation.

In our proposal, this post-editing step is performed in an OL framework. As a result, the system will learn from the corrections made by the user, and will use them to train statistical models on-the-fly. These models will be applied to automatically post-edit the next segments, so that once the user has amended a mistake, he or she will not need to amend it again if the mistake appears in the following segments to revise. Since MT post-editing is a repetitive task, OL models can ease this step by automatically amending those repetitive errors. Additionally, in the scenario without resources, OL will not be used as an APE system but rather as a regular SMT system incorporating OL initialized with empty models.

It should be noted that in the two scenarios considered above, namely, APE with OL and regular SMT with OL, the OL module will start with empty models. In spite of the fact that the performance of OL will be initially low due to the lack of training samples, empirical results shown in Section 5.3 clearly reflect the ability of OL to learn without previously existing training data. For instance, it is shown that an SMT system with online learning starting from empty models is able to obtain better performance than that obtained using a regular SMT system trained from out-of-domain models when translating test documents belonging to different domains and language pairs.

Regarding the software used to carry out the experiments presented in this paper, both the APE module and the regular SMT system incorporating OL have been implemented by means of the Thot toolkit (Ortiz-Martínez et al., 2005).

## 5. Experiments

We have performed some experiments to assess the previously described techniques. In this section, we show the different corpora chosen, the experimental framework, and the results achieved by each of the systems.

### 5.1. Corpora

We have worked with three different corpora to perform the experiments with our technique. We have tried to cover different domains and languages, in order to assess the proposed system in different scenarios. Table 1 shows some statistics of these corpora.

First, we have chosen the English and Spanish versions of the *EMEA* corpus. This is a publicly available medical corpus, formed by documents from the *European Medicines Agency* (Tiedemann, 2009). This is an interesting corpus because it was also chosen by some previous related work (Simard and Foster, 2013), where its suitability for the considered techniiques was proven due to the corpus document-like discourse units, and the technical and specialized nature of the texts. However, our results are not straightaway comparable to previous works, because they do not explain the employed partitions, and the involved languages differ.



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

Second, we wanted to choose a non-public corpus, due to the fact that some of the web-based translation applications against which we wanted to compare could have incorporated those public corpora into their training material. Thus, we chose the English and Spanish versions of the *Xerox* corpus. This corpus was compiled in the European project TransType2 (Esteban et al., 2004) from printer manuals provided by Xerox, one of the project partners.

Third, in order to cover different domains and languages, we chose the *i3media* corpus. It is a large corpus composed by newspaper articles in Catalan and Spanish (Lagarda et al., 2010). Its sentences are written in a richer language (with a more extensive vocabulary), covering several domains. In addition, Catalan and Spanish are similar languages, so we will be able to explore how our proposals behave in this situation.

Finally, as we have explained in previous sections, we have taken an out-of-domain corpus to assess the proposed techniques. For this purpose, we have chosen the seventh version of Europarl (Koehn, 2005).

We have divided each corpus in training, development and test sets. In order to preserve some of the context, we have grouped the sentences in blocks before randomly shuffling them and dividing them among the partitions. For clarity reasons, we only show figures for training and test partitions. Development numbers are similar to those from the test set, and it is used to tune the statistical model parameters of the offline SMT systems, where applicable.

Out-of-vocabulary (OOV) and perplexity figures give an idea of the complexity of each corpus. Perplexity is defined as the geometric average probability assigned by the model to each word in the test set (see for example Chen and Goodman (1996) for a formal definition). In Table 1, the perplexity has been computed for a 5-gram language model estimated from each (in-domain) training set using the IRSTLM toolkit (Federico et al., 2008). It is interesting to note how the Xerox task has radically different perplexities in each language. This can be related to the original language of the corpus (English) and its less complex translations into Spanish (Lembersky et al., 2012). Also, note that perplexities increase substantially with each corpus used, indicating that the latter are more complex than the former, and thus, the texts are richer.

Finally, Table 1 also shows the repetition rate (Bertoldi et al., 2013) for each test set. The repetition rate provides a quantitative measure of the degree of document-internal repetition for a given corpus. For this purpose, this measure looks at the rate of non-singleton n-grams contained in a given text. More specifically, the rates of non-singleton n-grams from $n = 1$–4 are calculated and geometrically averaged, using a sliding window of 1000 words to make the rates comparable across different sized corpora. Here we use a slightly modified version in which the sliding window calculation is removed, since in real translation scenarios, the text to be translated is available beforehand and should be completely translated. Our modified version of the repetition rate, RR', is defined as follows:

$$\mathrm{RR}'(\mathcal{I}) = \left( \prod_{n=1}^{4} \frac{|\mathcal{I}_{n,1+}| - |\mathcal{I}_{n,1}|}{|\mathcal{I}_{n,1+}|} \right)^{1/4} \tag{10}$$

where $|\cdot|$ represents the length of a given set, $\mathcal{I}_{n,1+}$ represents the set of different $n$-grams contained in the in-domain corpus $\mathcal{I}$, and $\mathcal{I}_{n,1}$ represents the set of different $n$-grams occurring only once in $\mathcal{I}$.

A more repetitive task will take advantage of the OL techniques (Simard and Foster, 2013). Repetition rates in Table 1 show how EMEA is a rather repetitive corpus, which can boost OL techniques. On the opposite side, i3media presents a lower repetition rate. It is interesting to explain the reason behind the differences in the observed repetition rates for each corpus. For this purpose, we can take a look to the way in which each of them was created. As it was mentioned above, the three used corpora consist of a collection of documents. Each one of the documents define some sort of a *sub-domain* that may change radically in the subsequent document. Analyzing the frequencies for the different document lengths, we observed a huge difference in those of i3media, the corpus with the lowest repetition rate, with respect to the document lengths for the other two corpora. More specifically, we found that the median length of an i3media document was 8 sentences, with 95% of the documents being shorter than 30 sentences. These figures are much higher for the other corpora, which presented documents composed of hundreds or even thousands of sentences. Therefore, the lower length of the i3media documents caused a much more frequent domain drift than that observed for the other two corpora, greatly decreasing the repetition rate.

We think it is important to stress out that the necessity of having a certain degree of repetition, so as to obtain translation quality improvements, cannot be seen as a specific limitation of our proposal, but as a limitation of domain adaptation techniques in general. In batch learning scenarios, domain adaptation data should be representative of the document to be translated in order to observe gains in translation quality. In other words, to ensure the correct translation of a given n-gram contained in the test corpus, it is necessary that this n-gram or similar ones have been seen in the



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

domain adaptation data. When we operate in online learning scenarios, we still have the same requirement but now the training and translation stages are no longer separated. Therefore, measuring the repetition rate of a given document could be seen as the equivalent of evaluating the representativity of domain adaptation data in a batch learning setting. In addition to this, sufficiently high repetition rates for test documents are common according to the document-internal repetition property (Church and Gale, 1995).

### 5.2. Experimental framework

For each one of the corpora (EMEA, Xerox, and i3media), we have carried out the following experiments:

W/O RESOURCES We consider the case where there is no base MT translation system, so the user manually translates the source language sentences.

RBMT Translation of the test set by means of the base RBMT systems. Depending on the languages pairs, we have employed *Apertium* and *PAHOMTS*.

WEB Translation of the test set by means of the base web systems. We have employed the translation engines from *Google*, *Bing*, and *Yandex*.

SMT In our premise, there are no in-domain corpora to train SMT models. To simulate this scenario, we have chosen an out-of-domain corpus (Europarl) to train the statistical models and translate each corpus test by means of Moses. All the Moses models have been trained employing the standard procedure, using a development set from the corpus to optimize their weights with MERT. These results are labelled in the plots as *Moses-OOD*.

In addition, as an optimistic scenario where we have a big enough in-domain corpus, we have trained SMT models with the corresponding training set of each corpus, and we have tuned them with their development sets. These results appear in the plots labelled as *Moses-ID*, and they are taken into account as an optimistic upper bound for these techniques, due to the fact that in the real scenario that we have considered, in-domain corpora are not frequent (Irvine et al., 2013). This is why we have tagged it as **ORACLE** in our plots. We have also tagged *Google* as an **ORACLE**, separating it from the rest of MT systems in those cases where we suspect that *Google* has probably been trained with in-domain corpora.

On the other hand, user corrections in the OL framework have been simulated with the reference translations of the test sets. After an empty initialization of the OL models, the process would be as follows for each test sentence:

1. The system provides the translation of the sentence (in the first sentence of each corpus, as the models are empty, it will propose the translation of the baseline system).
2. The user corrects the proposed translation (in our simulation, we use the reference translation directly, as a simulation of the user corrections).
3. The system learns from the corrections in order to enrich its models, which will be used to translate the next sentence.

In APE, this process takes, as source sentences, the translations given by the previous MT system.

We have automatically evaluated each system by means of the BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002) score, that has been extensively used in the SMT literature to measure the similarity between to texts. More specifically, the BLEU score computes the geometric mean of the precision of n-grams of various lengths between a hypothesis and a set of reference translations multiplied by a factor BP(·) that penalizes short sentences:

$$\mathrm{BLEU} = \mathrm{BP}(\cdot)\exp\left(\sum_{n=1}^{N}\frac{\log p_n}{N}\right)$$

where $p_n$ denotes the precision of n-grams in the hypothesis translation. Typically, a value of $N=4$ is used. In our experiments, the BLEU score gives us an idea of the effort needed by the user to post-edit the proposed translation. In the case of the baseline results without resources, we have measured BLEU against the source text. The results indicate how similar are the source sentences with respect to the references. In addition to this, we have tested the statistical significance of these results computing confidence intervals (95%) according to the method explained by Koehn (2004).



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

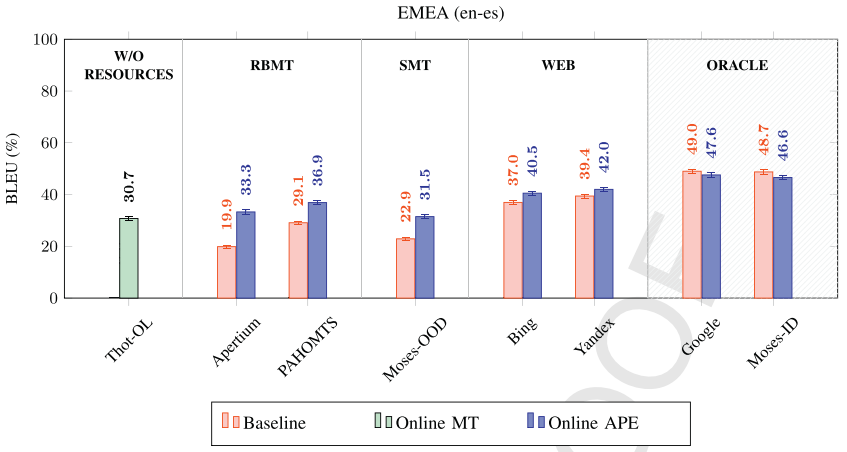



Fig. 3. EMEA English to Spanish. Comparison between each MT system translations and their online post-editings. Online learning translation with Thot is considered as baseline system. Confidence intervals at 95%.

### 5.3. Results

This section is devoted to the study and analysis of the experimental results for the different proposed corpora.

#### 5.3.1. EMEA

Fig. 3 compares the BLEU of the baseline systems with respect to their APE counterparts for the EMEA corpus from English to Spanish. In general, it can be observed how online APE outperforms significantly its baseline system translations. This means that OL models are able to learn from user corrections, and apply them successfully to post-edit the following sentences by adapting the baseline translation to the specific domain of the task.

Additionally, we can see some interesting results. OL improvements are more important when post-editing translations from systems with poorer translations (*Apertium*, *PAHOMTS*, and *Moses-OOD*), but are also very noticeable for the higher baseline **WEB** systems. With the **ORACLE** systems (*Google* and *Moses-ID*), post-editing is not able to improve translation quality. Although online APE systems perform a bit worse for them, differences are not statistically significant, i.e., the error bars overlap. Moreover, we should note that when SMT models are trained with out-of-domain corpora, their translations are much poorer than with in-domain data, as we expected. That is why in-domain training data is so critical for customizing SMT engines.

The extremely good *Google* baseline results could be explained by the fact that EMEA is a public corpus, and probably *Google* has incorporated the corpus to train its models. In fact, if we compare *Google* results with those achieved by *Moses-ID* (which is trained with the EMEA corpus), we see that they are quite similar. Moreover, the rest of web-based translation applications (*Bing*, *Yandex*) perform worse than *Google*, whereas in other corpora they perform more similarly. This confirms our suspicions to some extent. Also, both *Bing* and *Yandex* base results are improved when applying our technique pointing out that these **WEB** system do not poses specific domain information.

In addition, a standard online MT system is represented in the **W/O RESOURCES** experiment. Online MT assumes an OL SMT system that is initialized with empty models. By default, words that are unknown are left as they are in the source sentence. Thus, for the first sentences, the user has to post-edit the source sentences. However, as the system learns from the correct translations, the post-editing task is more similar to a classical SMT post-editing task. Online APE systems are also initialized with empty models but the output from the baseline system is copied into the target sentence instead of the source sentences. In this way, BLEU scores for the first sentences are expected to be higher for online APE systems that for online MT systems.

From the results, we can conclude that online APE systems outperform the online MT baseline. This implies that OL takes advantage of the translations given by the baseline MT systems more than when translating directly from the source language. Furthermore, we can see that online MT is better than the **RBMT** and **SMT** baseline because it is able to learn properly from the user translations. However, it cannot reach the quality of **WEB** systems, which are supposedly trained with huge amounts of data.



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

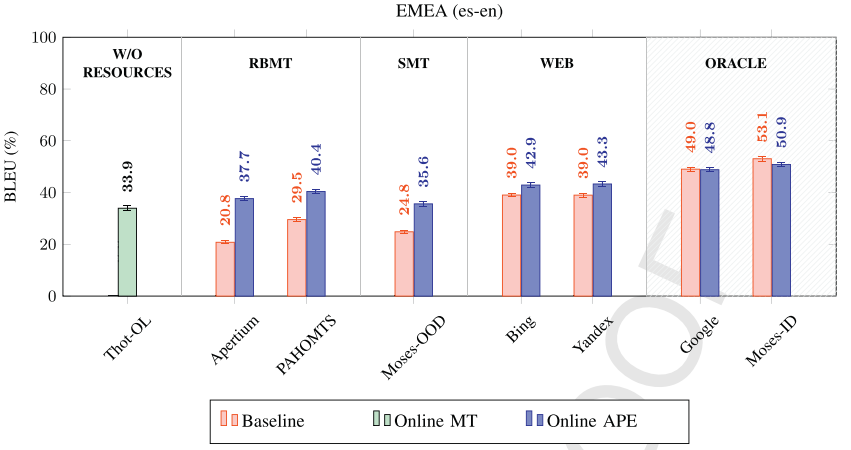



Fig. 4. EMEA Spanish to English. Comparison between each MT system translations and their online post-editings. Online learning translation with Thot is considered as baseline system. Confidence intervals at 95%.

With respect to the Spanish–English experiments, we can observe that Fig. 4 is essentially the same as Fig. 3. Therefore, we can conclude that, in this case, language direction does not pose additional problems.

Moreover, it should be noted that these results are obtained with the same corpus than in Simard and Foster (2013). However, our results are not directly comparable to those from Simard and Foster (2013), because our partition and languages are not the same as theirs. Nevertheless, we can see a similar behavior. When analyzing the out-of-domain SMT experiments, our method achieves around +[8.6,10.8] BLEU points compared to +[6.6,7.4] BLEU points in Simard and Foster (2013).

### 5.3.2. *Xerox*

As we have seen with the EMEA results, *Google* results were biased under the suspicion that *Google* probably included the EMEA corpus into their training material. Thus, we decided to repeat our experiments with a corpus that (hopefully) was not included in *Google* training.

Fig. 5 shows how *Google* performs much worse with respect to *Moses-ID*. In fact, it has not been considered as **ORACLE**. However, *Google* is still able to outperform their **WEB** counterparts. In addition, it is the only system, apart from Moses, where our technique, although achieving some improvements, is not able to achieve statistically significant results. In general, these results are quite consistent with those from EMEA. In almost all cases, our technique can improve the base translations with statistical significance.

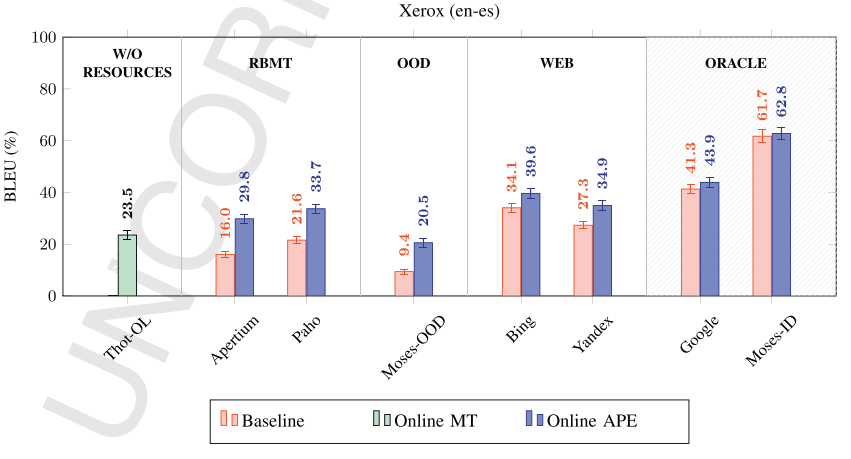



Fig. 5. Xerox English to Spanish. Comparison between each MT system translations and their online post-editings. Online learning translation with Thot is considered as baseline system. Confidence intervals at 95%.



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

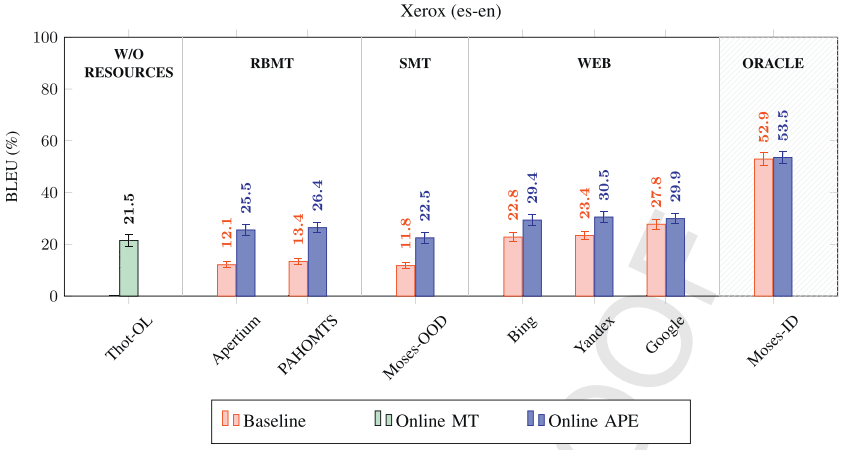



Fig. 6. Xerox Spanish to English. Comparison between each MT system translations and their online post-editings. Online learning translation with Thot is considered as baseline system. Confidence intervals at 95%.

When moving to the Spanish to English version of Xerox (Fig. 6), we can see that the improvements of our technique are coherent with those from English to Spanish. Nevertheless, BLEU scores are smaller here. This can be because the original corpus was generated translating from English to Spanish. As shown in Table 1, English perplexity is greater than the Spanish one, which points out that Spanish was translated in a, somehow, simplified version of Spanish. That is, the original Xerox manuals were written in English and after translated to Spanish, part of the source language complexity was lost (Lembersky et al., 2012).

### 5.3.3. i3media

Finally, we wanted to apply our technique to a different pair of languages. We chose the i3media corpus, which translates newspapers articles between Spanish and Catalan. Figs. 7 and 8 show how, in contrast to the previous corpora, our technique is not able to improve the baseline systems translations. Due to the fact that both languages are very similar in lexic and syntax, *Apertium*, *Google* and *Moses* achieve a very good baseline translations in both directions, suggesting that it is possible that OL APE does more harm than help.

However, even for *Bing* and *Yandex*, whose base BLEU are smaller, cannot take advantage of our technique, which worsen their initial scores. In the latter cases, it is even preferable to directly translate from the source language in an OL framework.

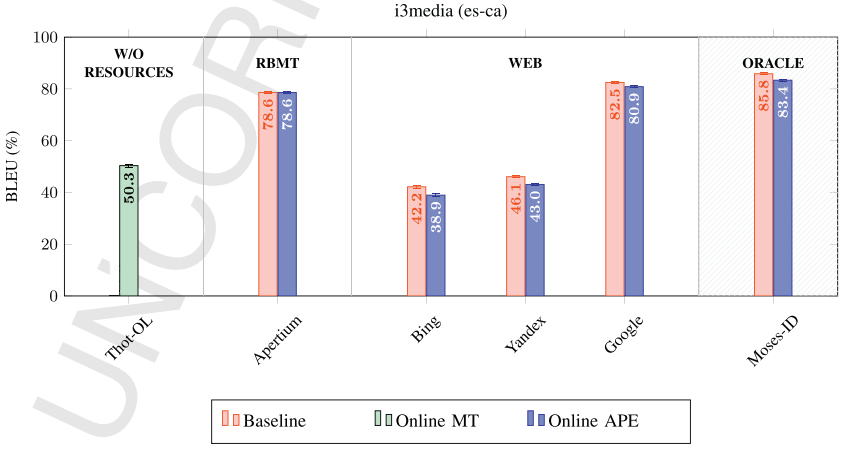



Fig. 7. i3media Spanish to Catalan. Comparison between each MT system translations and their online post-editings. Online learning translation with Thot is considered as baseline system. Confidence intervals at 95%.



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

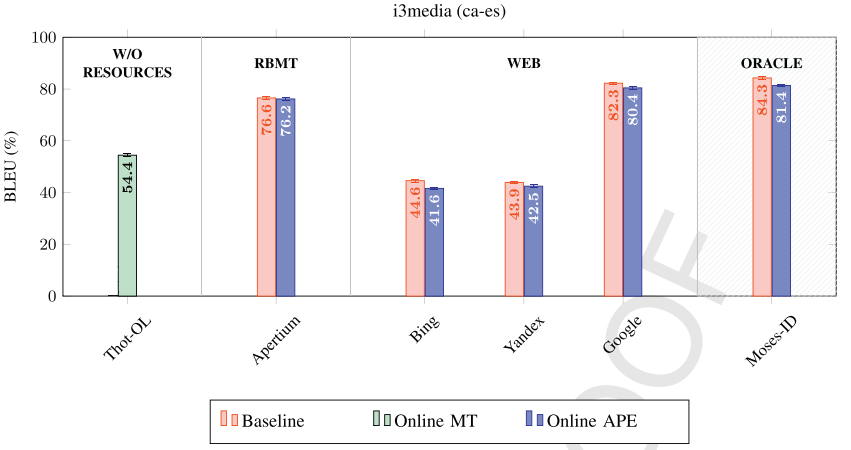



Fig. 8. i3media Catalan to Spanish. Comparison between each MT system translations and their online post-editings. Online learning translation with Thot is considered as baseline system. Confidence intervals at 95%.

In order to explain this behavior, we have to take into account that the n-gram repetition rate in the i3media corpus is much lower than in Xerox or EMEA (see Table 1). As it was already commented in Section 5.1, OL techniques are expected to perform better in repetitive tasks, because they will learn to correct those mistakes appearing several times in the test set. For instance, the first time that a mistake occurs, it will be corrected by the post-editor. If this mistake appears again in a posterior sentence, OL models will correct it automatically. This is probably why OL can improve base systems results in EMEA and Xerox, but not in i3media.

On slightly different note, we can see that the online MT baseline is very high in i3media, although the corpus repetition rate is low. This is due to the fact that the involved languages (Catalan and Spanish) are very similar (Lewis, 2009) in lexicon and word order in Spanish and Catalan are basically the same. In fact, if we did not change a word from the source test set, we would have more than 14 points of BLEU when comparing to the reference translation (in the plots, labelled as *W/O resources*). This is exactly what happens when a translation system faces an unknown source word, it outputs that word to let the user translate it (or leave it unaltered in the case of being invariable in both languages). In first iterations of OL, that is, when post-editing the first sentences, statistical models are empty. Thus, every source word will be unknown and copied to the output for these first sentences. As a result, when languages are similar, OL baseline BLEU will be higher.

### 5.3.4. *The dynamics of online learning APE*

Figs. A1, A2, A3, A4, A5 and A6 in Appendix A show the evolution of BLEU and repetition rate for a window of the latest 100 sentences, as the user completes the translation. Each plot displays four values. First, filled with brown color, the repetition rate shows the degree of repetition of the latest 100 sentences with respect to the previous sentences, as computed by Eq. (10). Second, the green line represents a the BLEU evolution of the Thot-OL system, i.e., plain online MT without previous resources. Its purpose is to provide a baseline where APE is not considered but only online learning. On the contrary, the blue line represents a baseline where a black box system is used, but also without APE nor online learning. Finally, the red line is the online APE system that we want to analyse. Note that we expect that the red line is always above the others. If the blue line is above the red line, that would mean that online learning is not being helpful to improve the black box system. On the other hand, if the green line is above the red line that would mean that online learning does not benefit from performing APE over the black box system. Also note that, as the online MT system and the repetition rate do not depend on the kind of black box system, they are always the same for the same corpus and language direction. In addition, the curves might seem very similar for the same corpus but for the reverse language direction since in reality the sub-domains present in each of the 100-sentence blocks are the same regardless of the language direction, hence obtaining similar repetition rates.

With respect to the results, it must be pointed out that the performance of the online MT system is quite correlated to the repetition rate of the latest 100 sentences. When there is a peak in the repetition rate, we can deduce that the latest block of 100 sentences presents a more homogeneous input than for the previous sentences. Then, the online



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

| | | |
|---|---|---|
| **source:** | | ' Determinación de marcadores del cartílago en pacientes con espondilitis anquilosante tratados con adalimumab ' . |
| *W/O resources*: | +0.6 | ' ~~Determinación de marcadores del cartílago en~~ pacientes con espondilitis anquilosante tratados con adalimumab ' . |
| *Apertium*: | −1.8 | ' ~~Determination of marcadores of the cartilage~~ in patient with espondilitis anquilosante treaties with adalimumab ' . |
| *PAHOMTS*: | +0.0 | Marker determination of the cartilage in patients with ankylosing spondylitis treated with adalimumab . |
| *Moses-OOD*: | −4.2 | ' ~~Liability of scores of cartílago in~~ patients with espondilitis anquilosante treaties with adalimumab ' . |
| *Bing*: | −7.3 | ' ~~Determination of markers of cartilage in patients with ankylosing spondylitis treated with~~/regulations . Evaluation ... |
| *Yandex*: | −0.7 | ' ~~Determination of markers of the cartilage~~ in patients with ankylosing spondylitis treated with adalimumab ' . |
| *Google*: | −8.4 | ' ~~Determination of cartilage markers in patients~~ with ankylosing spondylitis treated with adalimumab . ' |
| *Moses-ID*: | −8.4 | ' ~~Creatine markers of cartilage in patients~~ with ankylosing spondylitis treated with adalimumab . ' |
| **reference:** | | ' Evaluation of the articular cartilage biomarkers in patients with Ankylosing Spondylitis receiving Adalimumab ' |

(a) Sentence 3 in the EMEA (es-en) corpus.

| | | |
|---|---|---|
| **source:** | | Guía del usuario de Servicios de exploración de red de CentreWare xi |
| *W/O resources*: | +21.4 | ~~Guía~~/The del usuario ~~de Servicios de exploración de red de CentreWare xi~~/Services Installation Guide xi CentreWare ... |
| *Apertium*: | +14.3 | ^Services Installation Guide ~~of~~ the user ~~of Services of exploration of network of~~/Network Scanning CentreWare ~~xi~~ |
| *PAHOMTS*: | +6.9 | Service user ' s manual of ~~exploration of network of~~/Scanning Services Installation CentreWare ~~xi~~ |
| *Moses-OOD*: | +6.0 | handbook user Services ~~exploration network~~/Scanning The CentreWare xi |
| *Bing*: | +1.3 | The CentreWare network ~~scan~~/Scanning service user guide xi |
| *Yandex*: | +4.9 | User 's ~~guide exploration Services network~~/Guide Scanning The CentreWare ~~xi~~ |
| *Google*: | −12.1 | User Services ^The CentreWare ~~Network~~/Guide Scanning ~~Guide xi~~ |
| *Moses-ID*: | −28.0 | User Guide of CentreWare ~~Network~~/xi Scanning ~~Services xi~~ |
| **reference:** | | The CentreWare Network Scanning Services User Guide xi |

(b) Sentence 7 in the Xerox (es-en) corpus.

| | | |
|---|---|---|
| **source:** | | I jo somreia i m ' ho creia . |
| *W/O resources*: | +3.6 | I jo somreia ~~i m '~~/y ho creia . |
| *Apertium*: | −40.3 | Y yo sonreía y ~~me~~ lo creía . |
| *Bing*: | +0.0 | Sonreí y me lo creí . |
| *Yandex*: | −1.3 | Y yo sonrió ~~y m '~~/' Toda esa creí . |
| *Google*: | +0.0 | Y yo sonreía im ' lo creía . |
| *Moses-ID*: | −40.3 | Y yo sonreía y ~~me~~ lo creía . |
| **reference:** | | Y yo sonreía y me lo creía . |

(c) Sentence 8 in the i3media (ca-es) corpus.

Fig. 9. First changes of the Online APE system for various corpora. Positive BLEU increases by using Online APE are shown in **green** , whereas **red** indicates the contrary. Additionally, Online APE changes are shown as ~~deletions~~ from the baseline sentences, ^insertions in the baseline sentences and ~~substituted~~/substitution .

learning system is able to learn from this repetition to proportionally improve the BLEU scores. In fact, in EMEA the correlation is $r(4826) = 0.55$ for (en-es) and $r(4826) = 0.5$ for (es-en), whereas in Xerox it is $r(1124) = 0.62$ for (en-es) and $r(1124) = 0.74$ for (es-en), all with $p < .001$. This indicates that repetition rate is a good predictor for the BLEU in online MT system. However, i3media also behaves unexpectedly in this regard with $r(5099) = 0.12$ for (ca-es) and $r(5099) = 0.10$ for (es-ca) with $p < .001$, implying that repetition rates and BLEU are not correlated for this task. This confirms again our suspicion that i3media is a case with special properties for which our technique does not perform as expected.

In general, the dynamics are very consistent between all the systems for a given corpus. However Xerox dynamics presents some steep drops and rises in BLEU windows. Their main reason is the fact that the corpus is not shuffled. This corpus is composed of several blocks of sentences with slightly different sub-domains. These sudden changes in BLEU windows indicate that a new subdomain block has begun. To confirm this, we have calculated the percentage of those n-grams present in the test that were not seen in the training. For instance, we obtain that only a 9.8% of the n-grams appearing in test sentences from 400 to 600 do not appear in the training set. This is why BLEU for these sentences is quite high. On the other hand, when moving to the 800–1000 test block, a 16.8% of the n-grams do not appear in the training set. This explains the sudden drop around these sentences for Moses in-domain experiments.

Second, in almost all cases, **RBMT**+APE systems outperform their respective **RBMT** systems, except probably for the first sentences where the APE system has not received enough training data. For instance, Fig. 9 shows the first APE changes that appear in some of the corpora. In EMEA (Fig. 9a) and i3media (Fig. 9c) we can observe a series of meaningless deletions and substitutions provoked by phrases extracted in the previous sentences from uniformly aligned models, which were obviously wrong. On the contrary, in Xerox (Fig. 9b) almost any APE system has learned `Scanning` and `Services`, since they have been repeated in the previous 6 sentences. This repetition has allowed to build better alignment models from the beginning. The effect of the first sentences being wrongly post-edited with



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

| | | |
|---|---|---|
| **source:** | | singulair 4 mg comprimidos masticables |
| *W/O resources*: | +15.0 | ~~singulair~~/SINGULAIR 4 mg ~~comprimidos masticables~~/chewable tablets |
| *Apertium*: | +15.0 | ~~singulair~~/SINGULAIR 4 mg ~~compressed masticables~~/chewable tablets |
| *PAHOMTS*: | +20.8 | ~~singulair~~/SINGULAIR 4 ^mg chewable ~~compressed mg~~/tablets |
| *Moses-OOD*: | +0.0 | 4 ^chewable tablets ~~masticables~~ singulair mg |
| *Bing*: | +15.0 | ~~singulair~~/SINGULAIR 4 mg chewable tablets |
| *Yandex*: | +19.0 | ~~singulair 4 mg~~/SINGULAIR 4 mg chewable tablets |
| *Google*: | +15.0 | ~~singulair~~/SINGULAIR 4 mg chewable tablets |
| *Moses-ID*: | +15.0 | ~~singulair~~/SINGULAIR 4 mg chewable tablets |
| **reference:** | | SINGULAIR 4 mg chewable tablets |

(a) Sentence 3809 in the EMEA (es-en) corpus.

| | | |
|---|---|---|
| **source:** | | Instalación de la Utilidad de administración de fuentes |
| *W/O resources*: | +63.4 | ~~Instalación de la Utilidad de administración de fuentes~~/Using the Font Management Utility |
| *Apertium*: | +83.4 | ~~Installation of~~/Installing the ~~Utility of administration of sources~~/Font Management Utility |
| *PAHOMTS*: | +85.0 | ~~Installation of~~/Installing the ~~Usefulness of administration of sources~~/Font Management Utility |
| *Moses-OOD*: | +38.3 | ~~Deployment of Utilidad administration of sources .~~/Installing Use Font Management Utility |
| *Bing*: | +15.7 | ~~Installation sources management utility~~/Installing your Management Utility |
| *Yandex*: | +75.3 | ~~Installation of~~/Installing the ~~font management~~/Font Management Utility |
| *Google*: | +0.0 | Installing the Font Management Utility |
| *Moses-ID*: | +0.0 | Installing the Font Management Utility |
| **reference:** | | Installing the Font Management Utility |

(b) Sentence 98 in the Xerox (es-en) corpus.

| | | |
|---|---|---|
| **source:** | | Això és exactament el que va passar ahir . |
| *W/O resources*: | +84.3 | ~~Això és exactament el~~/Eso es exactamente lo que ~~va passar ahir~~/ocurrió ayer . |
| *Apertium*: | +72.2 | ~~Esto~~/Eso es exactamente ~~el~~/lo que ~~pasó~~/ocurrió ayer . |
| *Bing*: | +48.7 | ~~Esto~~/Eso es exactamente lo que ~~pasó~~/ocurrió ayer . |
| *Yandex*: | +48.7 | ~~Esto~~/Eso es exactamente lo que ~~sucedió~~/ocurrió ayer . |
| *Google*: | +48.7 | ~~Esto~~/Eso es exactamente lo que ~~pasó~~/ocurrió ayer . |
| *Moses-ID*: | +33.9 | Eso es exactamente lo que ~~sucedió~~/ocurrió ayer . |
| **reference:** | | Eso es exactamente lo que ocurrió ayer . |

(c) Sentence 3788 in the i3media (ca-es) corpus.

Fig. 10. Examples of Online APE corrections for various corpora. Positive BLEU increases by using Online APE are shown in **green** , whereas **red** indicates the contrary. Additionally, Online APE changes are shown as ~~deletions~~ from the baseline sentences, ^insertions in the baseline sentences and ~~substituted~~/substitution .

APE is especially noticeable for the i3media corpus since *Apertium* obtains a very good BLEU baseline around 78 points of BLEU. However, after more than 1000 sentences, *Apertium*+APE achieves small but persistent improvements over *Apertium* alone. That is, although our technique is not better than *Apertium* overall in i3media, it can improve the baseline after some sentences have been learned by the online APE system. In fact, if we use the baseline *Apertium* system for the first 1000 sentences until the *Apertium*+APE system has learned and then switch to it, we can reach 79.4 BLEU points for (es-ca) and 76.8 for (ca-es), which is better that either system alone. Hopefully, online APE systems begin to compensate the problem with uniformly initialized models soon enough so that the final BLEU score is usually better for online APE systems than for baseline systems. In Fig. 10 we can see how APE is able to amend some of the errors produced by the baseline systems in a way that BLEU scores increase. In particular, in Fig. 10a APE is able to learn the proper casing of `SINGULAIR` in the English documents but also it is able to substitute `chewable tablets` in **RBMT** systems and *Moses-OOD*. APE is also useful to amend casing in Fig. 10b. In addition, APE has been able to fix the lexical choices and word ordering of `Font Management Utility`, which should be consistent throughout all the printer manual. Finally, Fig. 10c shows how the APE system is able to modify a couple of lexical choices that are originally correct but probably the publisher prefers them to be translated as `Eso` and `ocurrió`.

Third, we should note that *Moses-OOD* behaves in a similar way to **RBMT** systems, probably due to the fact that it was trained only with an out-of-domain corpus. Similarly, most web-based translation systems are trained with corpora from multiple domains, but not likely with a corpus with the same exact domain of our test sets. In these cases,



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

the **SMT**+APE system is able to improve the baseline system in most of the cases, most notably where the peaks in repetition rate appear but the baseline system BLEU drops. On the other hand, in the valleys where we can perceive a drop in BLEU, the **SMT**+APE system tends to perform worse than the **SMT** system until the online learning algorithm adapts to the new sub-domain present in the 100-sentence window.

Additionally, *Moses-ID* and *Google*, which has been arguably trained with a bigger set of corpus and domains, are more difficult to beat by adapting the output with the APE system. This might be caused by these systems capturing already the nuances of the domain of the test set.

Finally, the **SMT**+APE models perform always worse than their **SMT** counterparts in the i3media corpus. As we have already explained, the language pairs are quite easy to learn for a **SMT** system, which is reflected by the high BLEU scores. In addition, this particular corpus presents a small repetition rate throughout all the test set. In the end, the APE system is not able to capture effectively any of the domain nuances more than the original **SMT** systems did.

## 6. Conclusions

In this work we have analyzed a real translation scenario, where a human expert needs to translate a document without having any similar translation memory, nor in-domain corpus to train SMT models. We have proposed the application of automatic post-editing in an online learning framework, similar to that shown in Simard and Foster (2013), but replacing the automatic post-editing module based on edit distance word alignments by a fully functional online SMT system (Ortiz-Martínez et al., 2010) that successfully removes any heuristic approximations introduced in previous works to update the model parameters.

We have assessed our technique with three corpora, namely, EMEA, Xerox and i3media, covering different domains and pairs of languages. Additionally, in the reported experiments we have combined our automatic post-editing module with three different kinds of base systems, including RBMT, Web and SMT systems. According to the obtained results, for the EMEA and Xerox corpora our system was able to significantly outperform the results obtained by all of the base systems. More specifically, improvements of up to 16, 11 and 7 BLEU points were obtained for the RBMT, Web and SMT systems respectively.

However, the proposed technique did not behave properly when facing our third corpus, i3media. As we have seen in our experiments, this is explained by the fact that the repetition rate of the n-grams appearing in the test set is lower than that observed for the other corpora used in the experimentation. From the results, we have proved the importance of n-gram repetitions to take advantage of OL techniques. This requirement has been previously suggested in other works (Simard and Foster, 2013). However, we think this is not an exclusive limitation of online learning but also of domain adaptation techniques in general (see Section 5.1 for a detailed explanation). On the other hand, in online learning scenarios, the document internal repetition property (Church and Gale, 1995) predicts a high probability of observing similar sentences composed of similar n-grams in a given document.

## Acknowledgements

Work partially supported by the European Union 7th Framework Programme (FP7/2007-2013) under the CasMaCat Project (Grant Agreement No. 287576), by Spanish MICINN under Grant TIN2012-31723, and by the Generalitat Valenciana under Grant ALMPR (Prometeo/2009/014).

## Appendix A. Dynamics of online learning APE

See Figs. A1–A6.



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

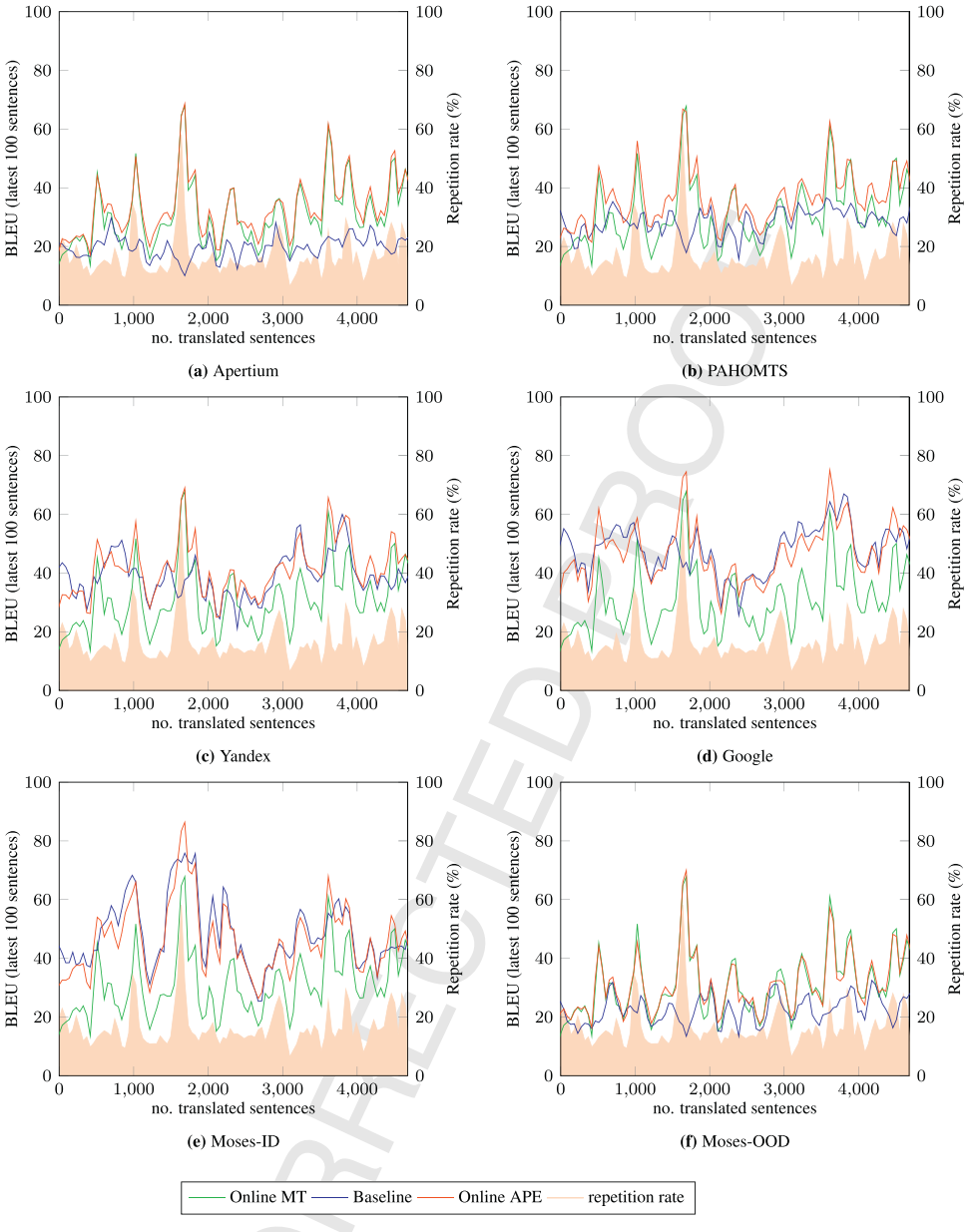



Fig. A1. Dynamics of online learning APE for the EMEA corpus (English to Spanish). The plots show the evolution of BLEU and repetition rate for a window of the latest 100 sentences, as the user completes the translation. The base system is compared to the APE system and a baseline online learning system.

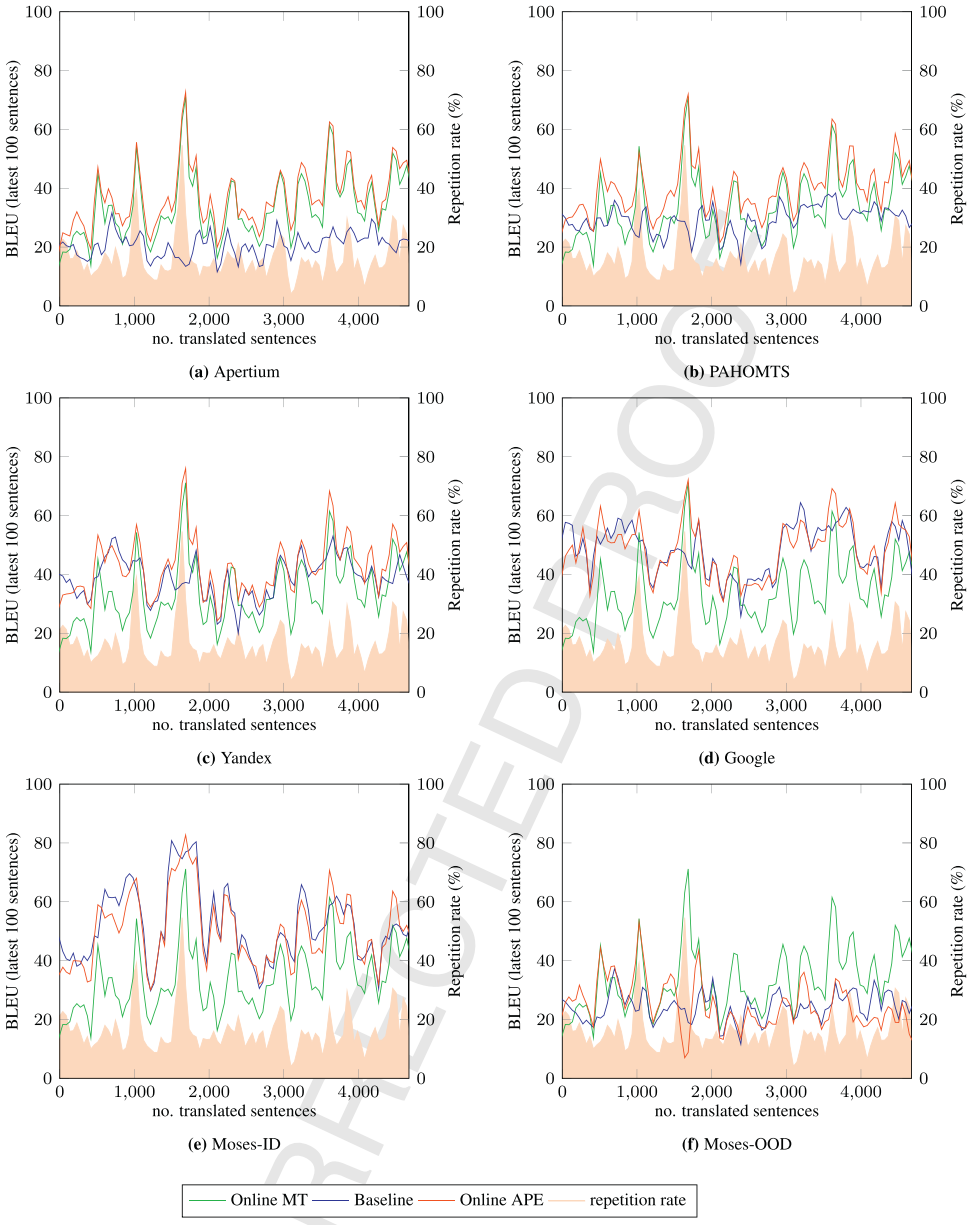



Fig. A2. Dynamics of online learning APE for the EMEA corpus (Spanish to English). The plots show the evolution of BLEU and repetition rate for a window of the latest 100 sentences, as the user completes the translation. The base system is compared to the APE system and a baseline online learning system.

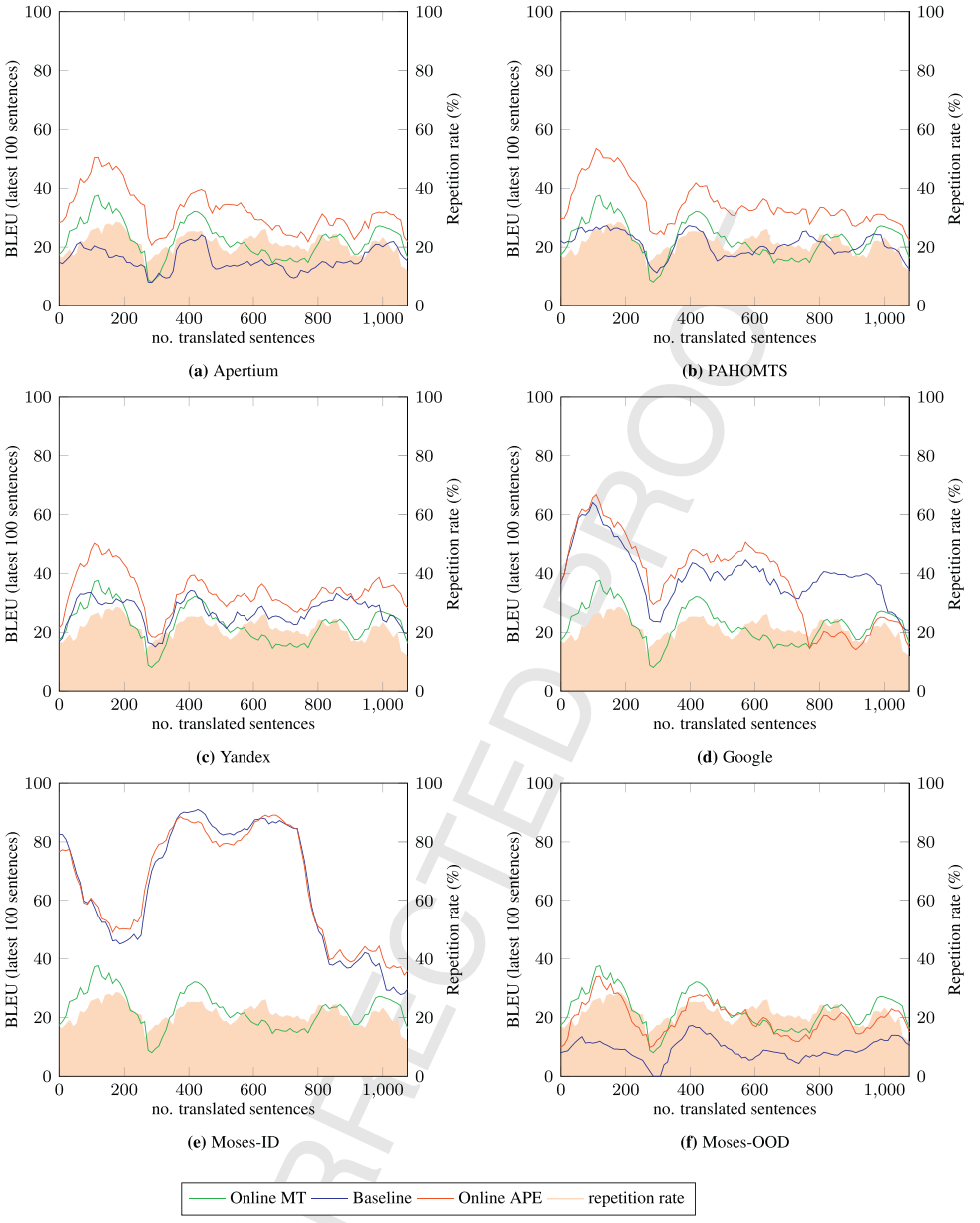



Fig. A3. Dynamics of online learning APE for the Xerox corpus (English to Spanish). The plots show the evolution of BLEU and repetition rate for a window of the latest 100 sentences, as the user completes the translation. The base system is compared to the APE system and a baseline online learning system.

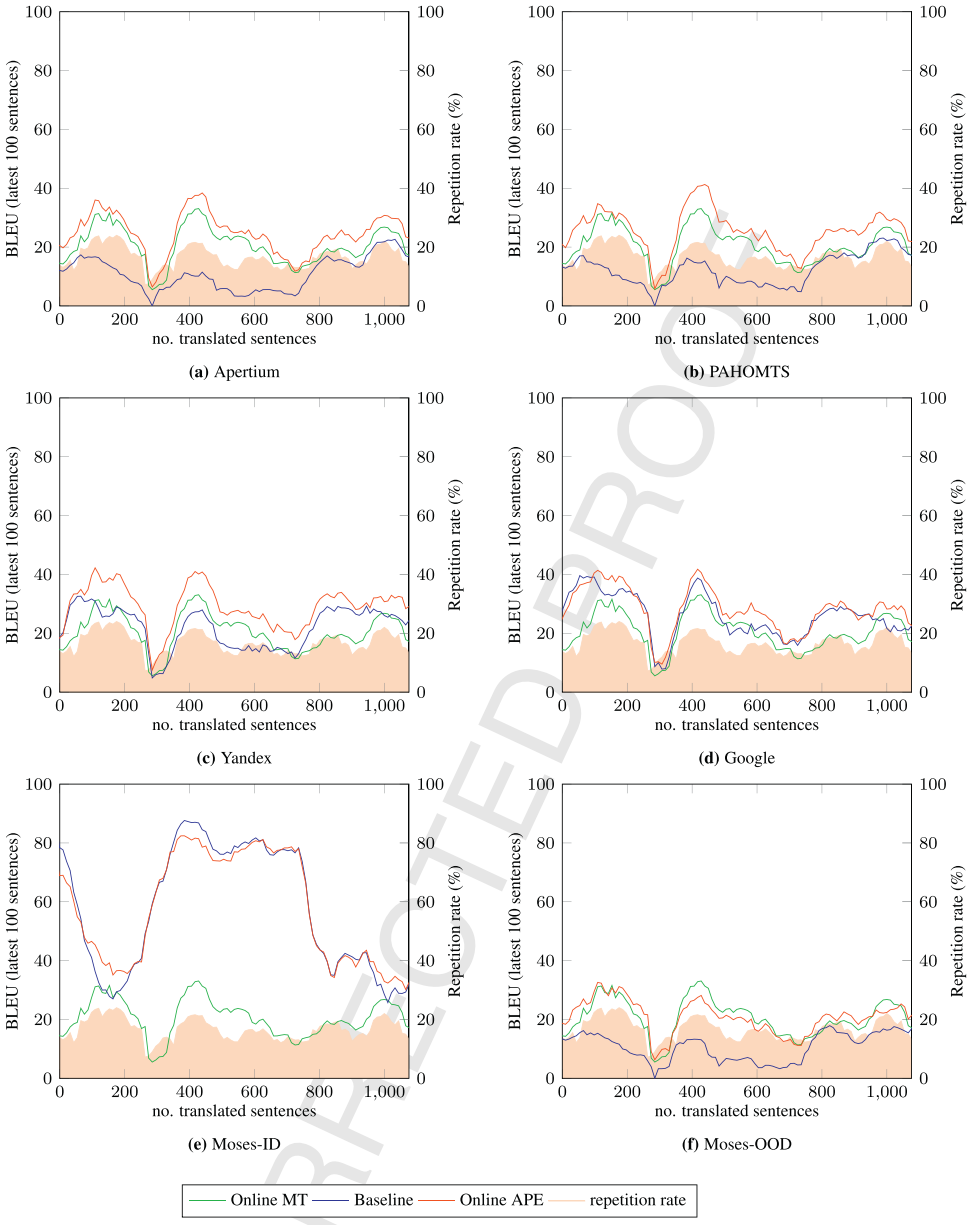
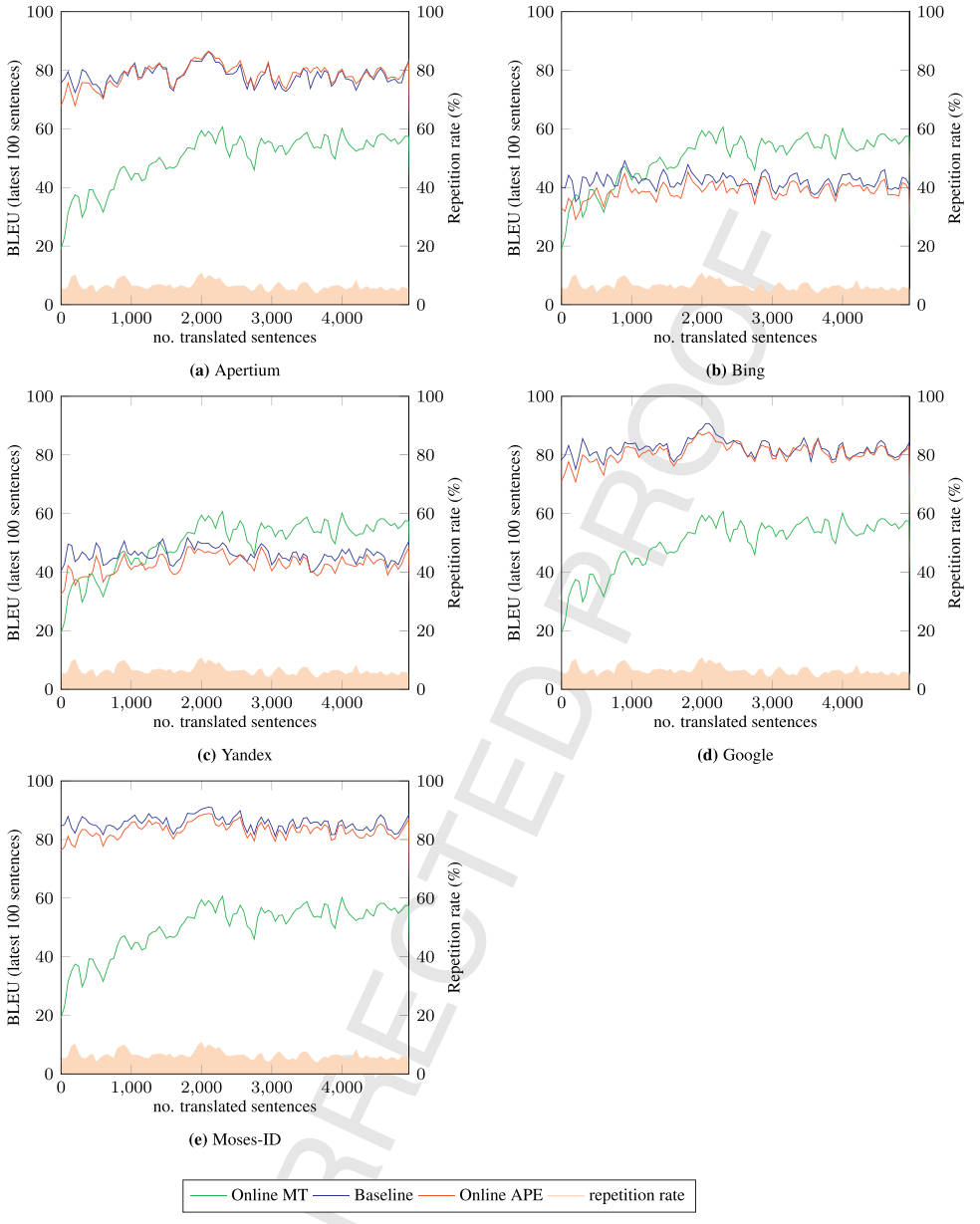



Fig. A4. Dynamics of online learning APE for the Xerox corpus (Spanish to English). The plots show the evolution of BLEU and repetition rate for a window of the latest 100 sentences, as the user completes the translation. The base system is compared to the APE system and a baseline online learning system.




**(a)** Apertium

**(b)** Bing

**(c)** Yandex

**(d)** Google

**(e)** Moses-ID

Fig. A5. Dynamics of online learning APE for the i3media corpus (Spanish to Catalan). The plots show the evolution of BLEU and repetition rate for a window of the latest 100 sentences, as the user completes the translation. The base system is compared to the APE system and a baseline online learning system.

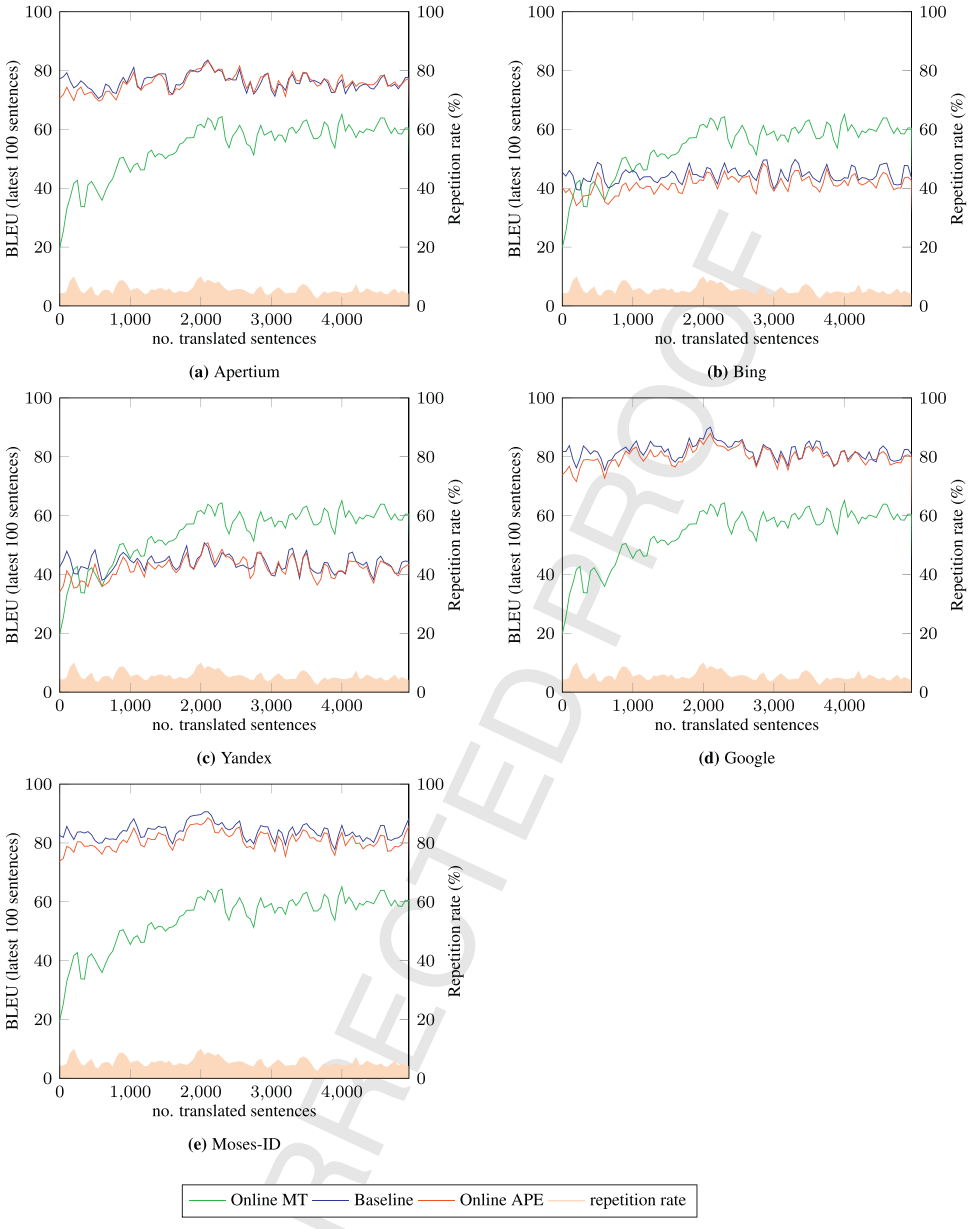



Fig. A6. Dynamics of online learning APE for the i3media corpus (Catalan to Spanish). The plots show the evolution of BLEU and repetition rate for a window of the latest 100 sentences, as the user completes the translation. The base system is compared to the APE system and a baseline online learning system.

## References

Alabau, V., Gonzlez-Rubio, J., Ortiz-Martínez, D., Sanchis-Trilles, G., Casacuberta, F., García-Martínez, M., Mesa-Lao, B., Cheung Petersen, D., Dragsted, B., Carl, M.,2014. Integrating online and active learning in a computer-assisted translation workbench. In: Proceedings of the First Workshop on Interactive and Adaptive Statistical Machine Translation. Association for Computational Linguistics.

Allen, J., Hogan, C., 2000. Toward the development of a post-editing module for machine translat ion raw output: a new productivity tool for processing controlled language. In: CLAW 2000: Third International Workshop on Controlled Language Applications, Seattle.

Béchara, H., Ma, Y., Genabith, G., 2011. Statistical post-editing for a statistical MT system. Xiamen, China, pp. 308–315.

Béchara, H., Rubino, R., He, Y., Ma, Y., van, J., Genabith, 2012. An evaluation of statistical post-editing systems applied to RBMT and SMT systems. In: COLING, pp. 5–230.

Bennett, W.S., Slocum, J., 1985. The LRC machine translation system. Comp. Linguist. 11 (2-3), 111–121.

Bertoldi, N., Cettolo, M., Federico, M., 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In: Machine Translation Summit, Nice, France, September.

Blain, F., Schwenk, H., Senellart, J., 2012. Incremental adaptation using translation information and post-editing analysis. In: International Workshop on Spoken Language Translation, Hong-Kong, China, pp. 234–241.

Brown, P., Della Pietra, S., Della Pietra, V., Mercer, S., 1993. The mathematics of statistical machine translation: parameter estimation. Comput. Linguist. 19 (2), 263–311.

Carpuat, M., Simard, M., 2012. The trouble with SMT consistency. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, WMT'12, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 442–449.

Cesa-Bianchi, N., Reverberi, G., Szedmak, S., 2008. Online learning algorithms for computer-assisted translation, Deliverable D4.2, SMART: Statistical Multilingual Analysis for Retrieval and Translation.

Chen, S.F., Goodman, J., 1996. An empirical study of smoothing techniques for language modeling. In: Joshi, A., Palmer, M. (Eds.), Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics, San Francisco, USA. Morgan Kaufmann Publishers, pp. 310–318.

Church, K.W., Gale, W.A., 1995. Poisson mixtures. Nat. Lang. Eng. 1, 163–190.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B: Methodol. 39 (1), 1–38.

Diaz, A., Labaka, G., Sarasola, K., 2008. Statistical post-editing: a valuable method in domain adaptation of RBMT systems for less-resourced languages. In: Mixing Approaches to Machine Translation, Donostia-San Sebastian, Spain, February, pp. 35–40.

Dove, C., Loskutova, O., de la Fuente, R., 2012. What's your pick: RBMT, SMT or hybrid?

Dugast, L., Senellart, J., Koehn, P., 2007. Statistical post-editing on SYSTRAN's rule-based translation system. In: Proceedings of the 2nd Workshop on SMT, Prague, Czech Republic. ACL, pp. 220–223.

Esteban, J., Lorenzo, J., Valderrábanos, A., Lapalme, G., 2004. Transtype2: an innovative computer-assisted translation system. In: Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions, ACLdemo'04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Federico, M., Bertoldi, N., Cettolo, M.,2008. Irstlm: an open source toolkit for handling large scale language models. In: INTERSPEECH. ISCA, pp. 1618–1621.

Forcada, M., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G., Tyers, F., 2011. Apertium: a free/open-source platform for rule-based machine translation. Mach. Trans. 25 (2), 127–144, Special Issue: Free/Open-Source Machine Translation.

Foster, G., Goutte, C., Kuhn, R., 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP'10, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 451–459.

Hardt, D., Elming, J., 2010. Incremental re-training for post-editing SMT. In: Proceedings of the 9th Annual Conference of the Association for Machine Translation in the Americas, Denver, CO, October.

Irvine, A., Morgan, J., Carpuat, M., Munteanu, D.S., 2013. Measuring machine translation errors in new domains. TACL 1, 429–440.

Isabelle, P., Goutte, C., Simard, M., 2007. Domain adaptation of MT systems through automatic post-editing. In: Proceedings of MT Summit XI, Copenhagen, Denmark, pp. 255–261.

Knight, K., Chander, I., 1994. Automated postediting of documents. In: Proceedings of AAAI.

Knuth, D.E., 1981. Seminumerical Algorithms, Volume 2 of the Art of Computer Programming, 2nd ed. Addison-Wesley, MA, USA.

Koehn, P., 2004. Statistical significance tests for machine translation evaluation. In: Lin, D., Wu, D. (Eds.), Proceedings of EMNLP 2004. Barcelona, Spain, July. Association for Computational Linguistics, pp. 388–395.

Koehn, P., 2005. Europarl: a parallel corpus for statistical machine translation. In: Conference Proceedings: The Tenth Machine Translation Summit, Phuket, Thailand. AAMT, pp. 79–86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., 2007. Moses: Open source toolkit for statistical machine translation. In: Proceedings of ACL, Prague, Czech Republic, pp. 177–180.

Koehn, P., Och, F., Marcu, D., 2003. Statistical phrase-based translation. In: Proceedings of Human Language Technologies: The 2003 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 48–54.

Koehn, P., Schroeder, J., June 2007. Experiments in domain adaptation for statistical machine translation. In: Proceedings of the 2nd ACL Workshop on SMT, Prague, Czech Republic, pp. 224–227.

Lagarda, A.L., Civera, J., Juan, A., Casacuberta, F., 2010. Interactive pattern recognition and human language technology for digital audiovisual content processing. In: Diethe, T., Cristianini, N., Shawe-Taylor, J. (Eds.), WAPA, Volume 11 of JMLR Proceedings. , pp. 103–110.

Lagarda, A.L., Alabau, V., Casacuberta, F., Silva, R., Díaz-de Liaño, E., 2009. Statistical post-editing of a rule-based machine translation system. In: Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT) 2009, pp. 217–220.

Lembersky, G., Ordan, N., Wintner, S., 2012. Language models for machine translation: original vs. translated texts. Comput. Linguist. 38 (4), 799–825.

Lewis, P., Summer Institute of Linguistics, 2009. Ethnologue: Languages of the World. SIL International, ISBN 9781556712166.

Lü, Y., Huang, J., Liu, Q., 2007. Improving statistical machine translation performance by training data selection and optimization. In: EMNLP-CoNLL, 34, pp. 3–350.

Neal, R., Hinton, G., 1999. A view of the EM algorithm that justifies incremental, sparse, and other variants., pp. 355–368, ISBN 0-262-60032-3.



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004

Nepveu, L., Lapalme, G., Langlais, P., Foster, G., 2004. Adaptive language and translation models for interactive machine translation. Barcelona, Spain, pp. 190–197.

Och, F., Ney, H., 2002. Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 295–302.

Oflazer, K., El-Kahlout, I.D., 2007. Exploring different representational units in English-to-Turkish statistical machine translation. In: Proceedings of the Second Workshop on Statistical Machine Translation, StatMT'07, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 25–32.

Ortiz-Martínez, D., García-Varea, I., Casacuberta, F., 2005. Thot: a toolkit to train phrase-based statistical translation models. In: Tenth Machine Translation.

Ortiz-Martínez, D., García-Varea, I., Casacuberta, F.,2010. Online learning for interactive statistical machine translation. In: HLT-NAACL. The Association for Computational Linguistics, pp. 546–554.

Ortiz-Martínez, D., Leiva, L.-A., Alabau, V., García-Varea, I., Casacuberta, F.,2011. An interactive machine translation system with online learning. In: Proceedings of the ACL-HLT 2011 System Demonstrations. Association for Computational Linguistics, pp. 68–73.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of ACL, Philadelphia, PA, USA, pp. 311–318.

Rubino, R., Huet, S., Lefèvre, F., Lenarés, G., 2012. Statistical post-editing of machine translation for domain adaptation. Proceedings of the European Association for Machine Translation (EAMT) 22, 1–228.

Silva, R., 2012. Post-editing integration in a translation agency workflow. In: International Workshop on Expertise in Translation and Post-editing Research and Application.

Simard, M., Foster, G., 2013. Pepr: Post-edit propagation using phrase-based statistical machine translation. In: Machine Translation Summit, Nice, France, September.

Simard, M., Goutte, C., Isabelle, P., 2007. Statistical phrase-based post-editing. In: Proceedings of NAACL-HLT2007, Rochester, NY. ACL, pp. 508–515.

Terumasa, E., 2007. Rule based machine translation combined with statistical post editor for Japanese to English patent translation. In: MT Summit XI Workshop on patent translation, Copenhagen, Denmark, pp. 13–18.

Tiedemann, J., 2009. News from OPUS: a collection of multilingual parallel corpora with tools and interfaces. In: Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (Eds.), Recent Advances in Natural Language Processing, vol. V. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, pp. 237–248.

Vasconcellos, M., León, M., April 1985. Spanam and engspan: Machine translation at the pan american health or ganization. Comput. Linguist. 11 (2-3), 122–136.

Vogel, S., Ney, H., Tillmann, C., 1996. HMM-based word alignment in statistical translation. In: Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark, pp. 836–841.

Wäeschle, K., Simianer, P., Bertoldi, N., Riezler, S., Federico, M., 2013. Generative and discriminative methods for online adaptation in SMT., pp. 11–18, Nice, France.

Zhao, B., Eck, M., Vogel, S., 2004. Language model adaptation for statistical machine translation via structured query models. In: COLING.



Please cite this article in press as: Lagarda, A.L., et al., Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.10.004