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