Pattern Recognition Letters 37 (2014) 124–134

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Cost-sensitive active learning for computer-assisted translation

Jesús González-Rubio [a,*], Francisco Casacuberta [b]

[a] Institut Tecnológic d'Informatica, Universitat Politècnica de València, Camino de Vera s/n, 46022 València, Spain
[b] D. Sistemes Informàtics i Computació, Universitat Politècnica de València, Camino de Vera s/n, 46022 València, Spain

ARTICLE INFO

Article history:
Available online 19 June 2013

Keywords:
Computer-assisted translation
Interactive machine translation
Active learning
Online learning

ABSTRACT

Machine translation technology is not perfect. To be successfully embedded in real-world applications, it must compensate for its imperfections by interacting intelligently with the user within a computer-assisted translation framework. The interactive–predictive paradigm, where both a statistical translation model and a human expert collaborate to generate the translation, has been shown to be an effective computer-assisted translation approach. However, the exhaustive supervision of all translations and the use of non-incremental translation models penalizes the productivity of conventional interactive–predictive systems.

We propose a cost-sensitive active learning framework for computer-assisted translation whose goal is to make the translation process as painless as possible. In contrast to conventional active learning scenarios, the proposed active learning framework is designed to minimize not only how many translations the user must supervise but also how difficult each translation is to supervise. To do that, we address the two potential drawbacks of the interactive–predictive translation paradigm. On the one hand, user effort is focused to those translations whose user supervision is considered more "informative", thus, maximizing the utility of each user interaction. On the other hand, we use a dynamic machine translation model that is continually updated with user feedback after deployment. We empirically validated each of the technical components in simulation and quantify the user effort saved. We conclude that both selective translation supervision and translation model updating lead to important user-effort reductions, and consequently to improved translation productivity.

1. Introduction

Machine translation (MT) is a fundamental technology that is emerging as a core component of natural language processing systems. A good example of multilingualism with high translation needs can be found in the European Union (EU) political institutions. According to EC (2009), the EU employs 1,750 full-time translators. Additionally, to cope with demand fluctuations, the EU uses external translation providers which generate approximately one fourth of its translation output. As a result, in 2008 the EU translation services translated more than 1,800,000 pages and spent about one billion Euros on translation and interpreting.

Besides being an expensive and time-consuming task, the problem with translation by human experts is that the demand for high-quality translation has been steadily increasing, to the point where there are just not enough qualified translators available today to satisfy it. This poses a high pressure on translation agencies that must decide how to invest their limited resources (budget, manpower, time, etc.) to generate translations of the maximum quality in the most efficient way.

To address this challenge, many translation agencies have focused their interest on MT technology. However, current state-of-the-art MT systems are still far from generating error-free translations (NIST, 2006; Lopez, 2008). Indeed, they usually require human experts to post-edit their automatic translations. This serial process prevents MT systems from taking advantage of the knowledge of the human experts, and the users cannot take advantage of the adaptive ability of MT systems.

An alternative way to utilize the existing MT technologies is to use them in collaboration with human translators within a computer-assisted translation (CAT) framework (Isabelle and Church, 1998). An important contribution to CAT technology was carried out during the TransType project (Foster et al., 1998; Langlais et al., 2000; Foster, 2002; Langlais and Lapalme, 2002). They proposed the interactive–predictive machine translation (IMT) framework where data-driven MT technologies are embedded within the translation environment. Following these ideas, Barrachina et al. (2009) proposed an innovative embedding where a fully-fledged statistical MT (SMT) system is used

* Corresponding author. Tel.: +34 96 387 70 69; fax: +34 96 387 72 39.
E-mail addresses: jegonzalez@iti.upv.es (J. González-Rubio), fcn@iti.upv.es (F. Casacuberta).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.06.007

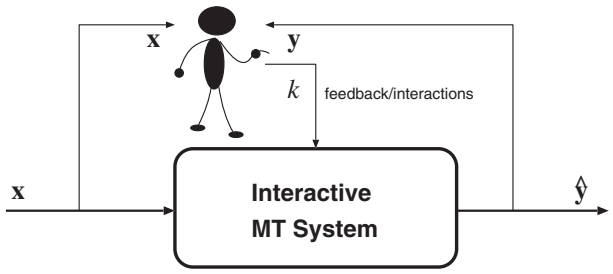

Fig. 1. Diagram of an interactive–predictive MT system. To translate a source sentence **x**, the user interacts with the system accepting or correcting the proposed translations **y**. User feedback k is used by the system to improve its suggestions.

to produce complete translations, or portions thereof, which can be accepted or amended by a human expert, see Fig. 1. Each corrected text segment is then used by the SMT system as additional information to achieve further, hopefully improved, translations.

Despite being an efficient CAT protocol, conventional IMT technology has two potential drawbacks. First, the user is required to supervise all the translations. Each translation supervision involves the user reading and understanding the proposed target language sentence, and deciding if it is an adequate translation of the source sentence. Even in the case of error-free translations, this process involves a non-negligible cognitive load. Second, conventional IMT systems consider static SMT models. This implies that after being corrected the system may repeat its errors, and the user will be justifiably disappointed.

We propose a cost-sensitive active learning (AL) (Angluin, 1988; Atlas et al., 1990; Cohn et al., 1994; Lewis and Gale, 1994) framework for CAT where the IMT user-machine interaction protocol (Fig. 2) is used to efficiently supervise automatic translations. Our goal is to make the translation process as efficient as possible. That is, we want to maximize the translation quality obtained per unit of user supervision effort. Note that this goal differs from the goal of traditional AL scenarios. While they minimize the number of manually-translated sentences to obtain a robust MT system, we aim at minimizing the number of corrective actions required to generate translations of a certain quality.

The proposed cost-sensitive AL framework boosts the productivity of IMT technology by addressing its two potential drawbacks. First, we do not require the user to exhaustively supervise all translations. Instead, we propose a selective interaction protocol where the user only supervises a subset of "informative" translations (González-Rubio et al., 2010). Additionally, we test several criteria to measure this "informativeness". Second, we replace the batch SMT model by an incremental SMT model (Ortiz-Martínez et al., 2010) that utilizes user feedback to continually update its parameters after deployment. The potential user effort reductions of our proposal are twofold. On the one hand, user effort is focused on those translations whose supervision is considered most "informative". Thus, we maximize the utility of each user interaction. On the other hand, the SMT model is continually updated with user feedback. Thus, the SMT model is able to learn new translations and to adapt its outputs to match the user's preferences which prevents the user from making repeatedly the same corrections.

The remainder of this article is organized as follows. First, we briefly describe the SMT approach to translation, and its application in the IMT framework (Section 2). Next, we present the proposed cost-sensitive AL framework for CAT (Section 3). Then, we show the results of experiments to evaluate our proposal (Section 4). Finally, we summarize the contributions of this article in Section 5.

source (**x**): Para ver la lista de recursos

desired translation ($\hat{\mathbf{y}}$): To view a listing of resources

interaction-0	$\mathbf{y}_p$	
	$\mathbf{y}_s$	*To view the resources list*
interaction-1	$\mathbf{y}_p$	To view
	k	**a**
	$\mathbf{y}_s$	*list of resources*
interaction-2	$\mathbf{y}_p$	To view a list
	k	**i**
	$\mathbf{y}_s$	*ng resources*
interaction-3	$\mathbf{y}_p$	To view a listing
	k	**o**
	$\mathbf{y}_s$	*f resources*
accept	$\mathbf{y}_p$	To view a listing of resources

Fig. 2. IMT session to translate a Spanish sentence into English. At interaction-0, the system suggests a translation ($\mathbf{y}_s$). At interaction-1, the user moves the mouse just before the first error and implicitly validates the first eight characters "To view " as a correct prefix ($\mathbf{y}_p$). Then, the user introduces a correction by pressing the **a** key (k). Lastly, the system suggests completing the translation from the user correction with "*list of resources*" (a new $\mathbf{y}_s$). At interaction 2, the user validates "*To view a list*" and introduces a correction **i** which is completed by the systems to form a new translation "*To view a listing of resources*". Interaction 3 is similar. Finally, the user accepts the current translation which is equal to the desired translation.