Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Computer Speech & Language xxx (2016) xxx-xxx
www.elsevier.com/locate/csl

Interactive neural machine translation

Álvaro Peris*, Miguel Domingo, Francisco Casacuberta

Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain

Received 18 April 2016; received in revised form 28 October 2016; accepted 2 December 2016

Abstract

Despite the promising results achieved in last years by statistical machine translation, and more precisely, by the neural machine translation systems, this technology is still not error-free. The outputs of a machine translation system must be corrected by a human agent in a post-editing phase. Interactive protocols foster a human−computer collaboration, in order to increase productivity. In this work, we integrate the neural machine translation into the interactive machine translation framework. Moreover, we propose new interactivity protocols, in order to provide the user an enhanced experience and a higher productivity. Results obtained over a simulated benchmark show that interactive neural systems can significantly improve the classical phrase-based approach in an interactive-predictive machine translation scenario.

Keywords: Neural machine translation; Interactive-predictive machine translation; Recurrent neural networks

1. Introduction

The statistical framework has allowed a breakthrough in machine translation (MT) and new systems provide admissible results for many tasks. However, in other scopes the quality of fully-automated systems is insufficient. In such cases, MT is used to obtain translation hypotheses, which must be supervised and corrected by a human agent in a post-editing (PE) stage. This working method is more productive than a completely manual translation, since the translator starts from an initial hypothesis that must be corrected. Nevertheless, this is a decoupled strategy in which computer and human agent work independently. Higher efficiency rates can be reached if human and system collaborate on a joint strategy. Seeking for this human−computer collaboration, Foster et al. (1997) introduced the so-called interactive-predictive MT (IMT), further developed by Alabau et al. (2013); Barrachina et al. (2009); Bender et al. (2005); Langlais and Lapalme (2002) and Macklovitch (2006).

This approach consists in an iterative prediction−correction process: each time the user corrects a word, the system reacts offering a new translation hypothesis, expected to be better than the previous one. In the basic IMT proposal, the user was constrained to follow a left-to-right protocol. Always was corrected the left-most wrong word from a translation hypothesis. This word, together with the previous ones, formed a *validated prefix*. At each iteration, the user validated a larger prefix and the system produced an appropriate suffix for completing the translation.

* Corresponding author.
E-mail address: lvapeab@prhlt.upv.es (Á. Peris), midobal@prhlt.upv.es (M. Domingo), fcn@prhlt.upv.es (F. Casacuberta).

http://dx.doi.org/10.1016/j.csl.2016.12.003
0885-2308/ 2016 Published by Elsevier Ltd.

IMT evolved during the years, introducing advances related to the generation of the new suffix (Azadi and Khadivi, 2015; Cai et al., 2013; Green et al., 2014b; Koehn et al., 2014; Ortiz-Martínez, 2011), and the possibility of suggesting more than one suffix (Koehn, 2010; Torregrosa et al., 2014). Other novelties came from profiting the use of the mouse, validating a prefix and suggesting a new suffix each time the user clicked into a position to type a word correction (Sanchis-Trilles et al., 2008). The addition of confidence measures aided the user to validate correct prefixes (González-Rubio et al., 2010a; 2010b). The use of online learning techniques was also studied, aiming to improve the system with the user feedback (Mathur et al., 2014; Nepveu et al., 2004; Ortiz-Martínez, 2016). Related to this, González-Rubio et al. (2012) explored the active learning protocol in an interactive post-edition stage. An interactive approach was also developed for hierarchical translation models (González-Rubio et al., 2013). Multimodal interaction integrated handwriting recognition/speech recognition into the IMT environment (Alabau et al., 2011; 2014). Green et al. (2014a) investigated the interactive use of translation memories. Nonetheless, the core of the user protocol remained the same in all these cited works. Recent works (González-Rubio et al., 2016) strove to overcome the prefix-based approach. One of the interactive protocols proposed in our work, relies on these latter ideas of breaking down the prefix constraint.

The prefix-based protocol suffered from three main issues: first, it was quite restrictive. The human translator was forced to always follow the left-to-right validation direction. This could be unnatural for the users or even inadequate in many cases. Second, the IMT system could produce worse suffixes, which also should be corrected by the user. Apart from increasing the human effort of the process, this introduced an annoying behaviour: the user had to correct words that were right in previous iterations, leading to user exasperation. The third issue was the computational cost of the (prefix-constrained) search for alternative hypotheses, which prevented the use of regular decoders. The increase of the computational power has alleviated this problem, allowing the use of more complex models and search strategies, in order to reach real-time generation of successive hypotheses.

Pursuing to overcome both first problems, in this work we propose an alternative protocol: when a hypothesis is generated, the user can select correct word sequences, called *segments*, from all over the sentence. These segments are considered to be valid and will remain in future iterations. The user can also correct wrong words, as in the classical approach. The system offers then an alternative hypothesis, that takes into account the corrected word together with the validated segments. Thus, correct parts of the hypothesis are kept during successive interactions, offering a more comfortable user experience and an increase in the productivity.

Up to now, the IMT approaches were based on discrete representations of words and sentences. Nevertheless, in the last years, continuous representations of words and sentences have gained much the attention of the natural language processing community. Distributed representations are richer than classical ones, yielding encouraging results. Although neural models were already applied to MT long ago (Castaño and Casacuberta, 1997), they finally took off recently and its use has dramatically increased. Bengio et al. (2003) proposed to project words into a distributed space and estimate the probabilities of a language model in such space. From here, continuous models have been used profusely in a wide range of tasks like language modelling (Mikolov et al., 2010; Schwenk, 2007; Sundermeyer et al., 2012), handwritten text recognition (Graves et al., 2009) or automatic speech recognition (Graves et al., 2013). In the MT field, neural models have been successfully introduced into the current statistical machine translation (SMT) pipeline, both in the phrase-based and hierarchical approaches (Devlin et al., 2014; Sundermeyer et al., 2014).

In addition to this, a neural approach to MT has been recently proposed (Cho et al., 2014; Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014). Neural machine translation (NMT) has emerged as one of the most promising technologies to tackle the MT problem. It is based on the use of neural networks for building end-to-end systems. The translation problem is addressed by a single, large neural network, which reads an input sentence and directly generates its translation. This is opposed to classical approaches to MT (e.g. Koehn et al., 2003), made up of multiple decoupled models. Most architectures are based on recurrent neural networks (RNN). In order to properly deal with long-term relationships, RNNs use gated units, such as long short-term memory (LSTM) units (Hochreiter and Schmidhuber, 1997) or gated recurrent units (GRU) (Cho et al., 2014).

There has been a significant effort for improving the NMT model. Thus, attention mechanisms were included to the model (Bahdanau et al., 2015; Luong et al., 2015b), allowing the model to focus on different parts of the input sentence. The out-of-vocabulary problem was tackled by Jean et al. (2015), Luong et al. (2015a) and Sennrich et al. (2016). Jean et al. (2015) also investigated the use of large target vocabularies. Gulcehre et al. (2015) included additional monolingual resources into the system. NMT at character-level has also obtained promising results (Chung et al., 2016; Costa-Jussà and Fonollosa, 2016; Ling et al., 2015).

Please cite this article as: &. Peris et al., Interactive neural machine translation, Computer Speech & Language (2016), http://dx.doi.org/10.1016/j.csl.2016.12.003