ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language xxx (2017) xxx-xxx

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Machine translation evaluation with neural networks

Q1 Francisco Guzmán, Shafiq Joty, Llúś Màrquez*, Preslav Nakov

ALT Research Group Qatar Computing Research Institute — HBKU, Qatar Foundation, Qatar

Received 1 June 2016; received in revised form 4 December 2016; accepted 7 December 2016

Abstract

We present a framework for machine translation evaluation using neural networks in a pairwise setting, where the goal is to select the better translation from a pair of hypotheses, given the reference translation. In this framework, lexical, syntactic and semantic information from the reference and the two hypotheses is embedded into compact distributed vector representations, and fed into a multi-layer neural network that models nonlinear interactions between each of the hypotheses and the reference, as well as between the two hypotheses. We experiment with the benchmark datasets from the WMT Metrics shared task, on which we obtain the best results published so far, with the basic network configuration. We also perform a series of experiments to analyze and understand the contribution of the different components of the network. We evaluate variants and extensions, including fine-tuning of the semantic embeddings, and sentence-based representations modeled with convolutional and recurrent neural networks. In summary, the proposed framework is flexible and generalizable, allows for efficient learning and scoring, and provides an MT evaluation metric that correlates with human judgments, and is on par with the state of the art.

Keywords: Machine translation; Reference-based MT evaluation; Deep neural networks; Distributed representation of texts; Textual similarity

1. Introduction

Automatic machine translation (MT) evaluation is a necessary step when developing or comparing MT systems. *Reference*-based MT evaluation, i.e., comparing the system output to one or more human reference translations, is the most common approach. Existing MT evaluation measures typically output an absolute quality score by computing the similarity between the machine- and the human-proposed translations. In the simplest case, the similarity is computed by counting word *n*-gram matches between the translation and the reference. This is the case of BLEU (Papineni et al., 2002), which has been the standard for MT evaluation for years. Nonetheless, more recent evaluation measures take into account various aspects of linguistic similarity and achieve better correlation with human judgments. For instance, synonymy and paraphrasing (Lavie and Denkowski, 2009), syntax (Giménez and Màrquez, 2007; Popović and Ney, 2007; Liu and Gildea, 2005), semantics (Giménez and Màrquez, 2007; Lo et al., 2012), and discourse (Comelles et al., 2010; Wong and Kit, 2012; Guzmán et al., 2014b; Joty et al., 2014).

* Corresponding author.
E-mail address: fguzman@qf.org.qa (F. Guzmán), sjoty@qf.org.qa (S. Joty), lmarquez@qf.org.qa (L. Màrquez), pnakov@qf.org.qa (P. Nakov).

http://dx.doi.org/10.1016/j.csl.2016.12.005
0885-2308/ 2017 Published by Elsevier Ltd.

The combination of all these aspects led to improved results in metric evaluation campaigns, such as the *WMT Metrics Shared Task* (Machacek and Bojar, 2014; Stanojević et al., 2015).

Having quality scores at the sentence level allows ranking alternative translations for a given source sentence. This is useful, for instance, for statistical machine translation (SMT) parameter tuning, for system comparison, and for assessing the progress during MT system development. The quality of automatic MT evaluation metrics is usually determined by computing their correlation with human judgments. To that end, quality rankings of alternative translations have been created by human judges. It is known that assigning an absolute score to a translation is a difficult task for humans. Hence, ranking-based evaluations, where judges are asked to rank the output of 2 to 5 systems, have been used in recent years, which has yielded much higher inter-annotator agreement (Callison-Burch et al., 2007).

These human quality judgments can be used to train automatic metrics. The supervised learning can be oriented to predict absolute scores, e.g., using regression (Albrecht and Hwa, 2008), or rankings (Duh, 2008; Song and Cohn, 2011). A particular case of the latter is used to learn in a pairwise setting, i.e., given a reference and two alternative translations (or hypotheses), the task is to decide which one is better. This setting emulates closely how human judges perform evaluation assessments in reality. From a machine learning perspective, the challenge is to learn, from a pair of hypotheses, which are the features that help to discriminate the better from the worse translation.

In previous work (Guzmán et al., 2014a), we presented a learning framework for this pairwise setting, based on preference kernels and support vector machines (SVM). We obtained promising results using a combination of syntactic and discourse-based structures. However, using convolution kernels over complex structures comes at a high computational cost both at training and at testing time because the use of kernels requires that the SVM operate in the much slower dual space. Thus, some simplification is needed to make it practical.

While there are some solutions in the kernel-based learning framework to alleviate the computational burden, we took a different direction and presented in Guzmán et al. (2015) the first neural network (NN) approach for MT evaluation, learning in the pairwise setting. The present article builds on that previous paper and explores some new additions while extending its analysis.

In the core NN model, lexical, syntactic and semantic information from the reference and the two hypotheses is compacted into relatively small distributed vector representations and fed into the input layer, together with a set of individual real-valued features coming from simple pre-existing MT evaluation metrics. A hidden layer, motivated by our intuitions on the pairwise ranking problem, is used to capture interactions between the relevant input components. Our evaluation results on the *WMT12 Metrics Shared Task* benchmark datasets (Callison-Burch et al., 2012) show high correlation with human judgments. These results clearly surpass Guzmán et al. (2014a) and are on par with the best results reported for this dataset, achieved by DiscoTK (Joty et al., 2014), which is a much heavier combination metric. Interestingly, we empirically show that the syntactic and semantic embeddings produce sizeable and cumulative gains in performance over a strong combination of pre-existing MT evaluation measures (BLEU, NIST, Meteor, and TER).

Another advantage of the proposed architecture is efficiency. Due to the vector-based compression of the linguistic structure and the relatively reduced size of the network, testing is fast, which would greatly facilitate the practical use of this approach in real MT evaluation and development.

In this paper, we broaden the discussion from Guzmán et al. (2015) by exploring two new model extensions, one oriented to fine-tuning the semantic embeddings on the task data, and the second to produce a sentence-level semantic representation of the input texts based on convolutional and recurrent neural networks. Better results could arguably be obtained by following these approaches, the tradeoff being substantial increase in complexity and reduction in efficiency/speed.

Additionally, we use the pairwise network to produce an absolute quality score when applied to a single input translation, i.e., as a standard MT evaluation metric. The pairwise setting is sufficient for most evaluation and MT development scenarios, and we claim that it should be preferred for the cases in which one has to compare a set of hypothesis translations to select the best one (ranking problem). However, one might also need to compare one's system to another system on a benchmark dataset, for which one knows the evaluation score but not the actual translations. In that case, the comparison requires the use of an evaluation metric that produces an absolute quality score for each system independently. As mentioned before, here we show how the network trained in the pairwise fashion can also be used to produce a high-quality MT evaluation metric over individual translations, which performs comparably to the state of the art both at the sentence and at the system levels.

Please cite this article as: F. Guzmán et al., Machine translation evaluation with neural networks, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2016.12.005