ARTICLE IN PRESS
JID: YCSLA [m3+;November 11, 2016;11:48]

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language xxx (2016) xxx-xxx

www.elsevier.com/locate/csl

Multi-way, multilingual neural machine translation☆

Orhan Firat[a,*], Kyunghyun Cho[b], Baskaran Sankaran[c], Fatos T. Yarman Vural[a], Yoshua Bengio[d,e]

[a] *Middle East Technical University, Turkey*
[b] *New York University, USA*
[c] *IBM T.J. Watson Research Center, USA*
[d] *University of Montreal, Canada*
[e] *CIFAR Senior Fellow, Canada*

Received 21 April 2016; received in revised form 23 August 2016; accepted 22 October 2016

Abstract

We propose multi-way, multilingual neural machine translation. The proposed approach enables a single neural translation model to translate between multiple languages, with a number of parameters that grows only linearly with the number of languages. This is made possible by having a single attention mechanism that is shared across all language pairs. We train the proposed multi-way, multilingual model on ten language pairs from WMT'15 simultaneously and observe clear performance improvements over models trained on only one language pair. We empirically evaluate the proposed model on low-resource language translation tasks. In particular, we observe that the proposed multilingual model outperforms strong conventional statistical machine translation systems on Turkish−English and Uzbek−English by incorporating the resources of other language pairs.

Keywords: Neural machine translation; Multi-lingual; Low resource translation

1. Introduction

1.1. Neural machine translation

It has been shown that a deep (recurrent) neural network can successfully learn a complex mapping between variable-length input and output sequences on its own. Some of the earlier successes in this task have, for instance, been handwriting recognition (Bottou et al., 1997; Graves et al., 2009) and speech recognition (Chorowski et al., 2015; Graves et al., 2006). More recently, a general framework of encoder-decoder networks has been found to be effective at learning this kind of sequence-to-sequence mapping by using two recurrent neural networks (Cho et al., 2014b; Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014).

A basic encoder-decoder network consists of two recurrent networks. The first network, called an encoder, maps an input sequence of variable length into a point in a continuous vector space, resulting in a fixed-dimensional

☆ This manuscript extends the earlier conference paper (Firat et al., 2016).

* Corresponding author.
E-mail addresses: orhan.firat.cs@gmail.com, orhan.firat@ceng.metu.edu.tr (O. Firat).

http://dx.doi.org/10.1016/j.csl.2016.10.006
0885-2308/ 2016 Published by Elsevier Ltd.

context vector. The other recurrent neural network, called a decoder, then generates a target sequence again of variable length starting from the context vector. This approach however has been found to be inefficient in Cho et al. (2014a) when handling long sentences, due to the difficulty in learning a complex mapping between an arbitrary long sentence and a single fixed-dimensional vector.

In Bahdanau et al. (2014), a remedy to this issue was proposed by incorporating an *attention mechanism* to the basic encoder−decoder network. The attention mechanism in the encoder−decoder network frees the network from having to map a sequence of arbitrary length to a single, fixed-dimensional vector. Since this attention mechanism was introduced to the encoder−decoder network for machine translation, neural machine translation, which is purely based on neural networks to perform full end-to-end translation, has become competitive with the existing phrase-based statistical machine translation in many language pairs (Gulcehre et al., 2015; Jean et al., 2015; Luong et al., 2015b).

1.2. Multilingual neural machine translation

Existing machine translation systems, mostly based on a phrase-based system or its variants, work by directly mapping a symbol or a subsequence of symbols in a source language to its corresponding symbol or subsequence in a target language. This kind of mapping is strictly specific to a given language *pair*, and it is not trivial to extend this mapping to work on multiple pairs of languages.

A system based on neural machine translation, on the other hand, can be decomposed into two modules. The encoder maps a source sentence into a continuous representation, either a fixed-dimensional vector in the case of the basic encoder−decoder network or a set of vectors in the case of attention-based encoder−decoder network. The decoder then generates a target translation based on this source representation. This makes it possible conceptually to build a system that maps a source sentence in any language to a common continuous representation space and decodes the representation into any of the target languages, allowing us to make a *multilingual machine translation* system.

This possibility is straightforward to implement and has been validated in the case of basic encoder−decoder networks (Luong et al., 2015a). It is however not so, in the case of the attention-based encoder−decoder network, as the attention mechanism, or originally called the alignment function in Bahdanau et al. (2014), is conceptually language pair-specific. In Dong et al. (2015), the authors cleverly avoided this issue of language pair-specific attention mechanism by considering only a one-to-many translation, where each target language decoder embedded its own attention mechanism. Also, we notice that both of these works have only evaluated their models on relatively small-scale tasks, making it difficult to assess whether multilingual neural machine translation can scale beyond low-resource language translation. In Zoph and Knight (2016), authors proposed multi-source encoder−decoder networks with attention for the case where two input sources are given at the same time to exploit multi-text (three-way parallel data). Their proposed attention mechanism consists of two parametric alignment functions, both of which attends two sources separately and then combines context vectors by concatenation. As this architecture still uses a separate attention module for each source-target pair and strictly requires the availability of multi-text, it is not clear how to extend it for the cases where only parallel text is given for multiple source and target pairs.

1.3. Multi-way, multilingual neural machine translation

In this paper, we first step back from the currently available multilingual neural translation systems proposed in Luong et al. (2015a), Dong et al. (2015), Zoph and Knight (2016) and ask the question of whether the attention mechanism can be shared across multiple language pairs. As an answer to this question, we propose an attention-based encoder−decoder network that admits a shared attention mechanism with multiple encoders and decoders. We use this model for all the experiments, which suggests that it is indeed possible to share an attention mechanism across multiple language pairs.

We train a single model with the proposed architecture on all the language pairs from the WMT'15; English, French, Czech, German, Russian and Finnish. We compare this multi-way, multilingual model against 10 single-pair models and show that they perform comparably to each other, while the number of parameters in the multilingual model is substantially smaller than that of all the single-pair models. This confirms that it is indeed possible to train a single attention-based network to perform multi-way translation.

Please cite this article as: O. Firat et al., Multi-way, multilingual neural machine translation, Computer Speech & Language (2016), http://dx.doi.org/10.1016/j.csl.2016.10.006