Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computer Speech & Language 00 (2017) 1–12

www.elsevier.com/locate/csl

Context-dependent word representation for neural machine translation[☆]

Heeyoul Choi[*,a], Kyunghyun Cho[b], Yoshua Bengio[c]

[a] *Handong Global University, Pohang, Republic of Korea*
[b] *New York University, New York, NY, USA*
[c] *University of Montreal, Montreal, QC, Canada*

Received 18 April 2016; received in revised form 23 November 2016; accepted 1 January 2017

Abstract

We first observe a potential weakness of continuous vector representations of symbols in neural machine translation. That is, the continuous vector representation, or a word embedding vector, of a symbol encodes multiple dimensions of similarity, equivalent to encoding more than one meaning of the word. This has the consequence that the encoder and decoder recurrent networks in neural machine translation need to spend substantial amount of their capacity in disambiguating source and target words based on the context which is defined by a source sentence. Based on this observation, in this paper we propose to contextualize the word embedding vectors using a nonlinear bag-of-words representation of the source sentence. Additionally, we propose to represent special tokens (such as numbers, proper nouns and acronyms) with typed symbols to facilitate translating those words that are not well-suited to be translated via continuous vectors. The experiments on En–Fr and En–De reveal that the proposed approaches of contextualization and symbolization improves the translation quality of neural machine translation systems significantly.

Keywords: Neural machine translation; Contextualization; Symbolization

1. Introduction

Neural machine translation is a recently proposed paradigm in machine translation, which is often entirely built as a single neural network (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015). The neural machine translation system, which often consists of an encoder and decoder, projects and manipulates a source sequence of discrete linguistic symbols (source sentence) in a continuous vector space, and decodes a target sequence of symbols (target sentence or translation.) This is contrary to the conventional machine translation systems, such as phrase-based statistical machine translation (Koehn et al., 2003), which work directly at the discrete symbol level.

In more detail, the first step of any neural machine translation system is to convert each atomic symbol into a corresponding continuous vector, which is often called as a word embedding. This step is done for each source word

[☆] This paper has been recommended for acceptance by Prof. R. K. Moore.

[*] Corresponding author.
E-mail address: heeyoul@gmail.com (H. Choi).

http://dx.doi.org/10.1016/j.csl.2017.01.007

independently of the other words and results in a source sequence of word embeddings. The encoder network, which is often implemented as a recurrent neural network, encodes this source sentence either into a single context vector (Sutskever et al., 2014; Cho et al., 2014b) or into a sequence of context vectors (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2015).

The decoder network, again a recurrent neural network, generates a translation word-by-word while being conditioned on the context representation of the source sentence. At each step of generation in the decoder, the internal hidden state of the decoder is updated first. The dot product between this hidden state and the output word embedding vector of each word in the target vocabulary is computed and normalized across all the target words, resulting in a probability distribution over the target vocabulary. A target word is selected based on this distribution, and the whole process is recursively repeated until the end-of-sequence symbol is generated.

Among different variants of neural machine translation, the attention-based approach (Bahdanau et al., 2015) has recently become *de facto* standard. It has been found to perform comparably to or better than the existing phrase-based statistical systems in many language pairs including En−Fr (Jean et al., 2015a), En−De (Jean et al., 2015a; 2015b; Luong et al., 2015a), En−Cs (Jean et al., 2015b), and En−Zh (Shen et al., 2015). Much of these recent improvements have been made by tackling, e.g., the attention mechanism (which is central to the attention-based neural translation system) and the computational issues arising from having a large target vocabulary.

Unlike these recent works, we focus on source- and target-side word embedding vectors in this paper. More specifically, we first notice that the transformation from and to high-dimensional word embedding vectors is done for each word largely independent of each other. We conjecture that only a few axes in this high-dimensional space are relevant given a source sentence and that we can remove much of the ambiguity in the choice of words by restricting, or turning off, most of the irrelevant dimensions. We propose to achieve this automated way to turn off some dimensions of word embeddings by *contextualizing* a word embedding vector.

In addition to the proposed contextualization of both source and target word embedding vectors, we propose to extend the unknown token replacement technique proposed in Luong et al. (2015b) to multiple token types. This extension, to which we refer as *symbolization*, introduces multiple meta-tokens such as the number token and the proper name token in addition to the unknown-word token. This symbolization effectively remaps rare tokens into more frequent meta-tokens and thereby results in improved translation quality.

We extensively evaluate the proposed contextualization and symbolization on two language pairs—En−Fr and En−De—with the attention-based neural machine translation model. The experiments reveal that the contextualization and symbolization each improves the translation quality by 2 BLEU scores, and together by 4 BLEU points on En−Fr. On En−De, they result in 1−2 BLEU score improvements each, and together 2.5 BLEU score increase.

2. Background: neural machine translation

In this section, we give a brief overview of neural machine translation. More specifically, we describe the attention-based neural machine translation (Bahdanau et al., 2015) which will be used in the experiments later. However, we note that the proposed contextualization and symbolization techniques are generally applicable to any other type of neural machine translation systems such as the sequence-to-sequence model (Sutskever et al., 2014).

The attention-based neural machine translation system computes a conditional distribution over translations given a source sentence $X=(w_1^x, w_2^x, \dots, w_T^x)$:

$$p\left(Y=(w_1^y, w_2^y, \dots, w_{T'}^y)|X\right).$$

This is done by a neural network that consists of an encoder, a decoder and the attention mechanism.

The encoder is often implemented as a bidirectional recurrent neural network that reads the source sentence word-by-word. Before being read by the encoder, each source word $w_t^x \in V$ is projected onto a continuous vector space:

$$\boldsymbol{x}_t = \boldsymbol{E}^x \mathbf{1}(w_t^x), \tag{1}$$

where $\mathbf{1}(w_t^x)$ is a one-hot vector defined as

$$\mathbf{1}(w_t^x)_j = \begin{cases} 1, & \text{if } j = w_t^x \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$