Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language xxx (2018) xxx-xxx

www.elsevier.com/locate/csl

Q1 Improving LSTM CRFs using character-based compositions for Korean named entity recognition

Q2 Seung-Hoon Na[a], Hyun Kim[b], Jinwoo Min[c], Kangil Kim*[,d]

[a] *Department of Computer Science, Chonbuk National University, South Korea*
[b] *Department of Creative IT Engineering, Pohang University of Science and Technology (POSTECH), South Korea*
[c] *Department of Computer Science, Chonbuk National University, South Korea*
[d] *Department of Computer Science, Konkuk University, South Korea*

Received 24 March 2017; received in revised form 27 January 2018; accepted 16 September 2018
Available online xxx

Abstract

Standard approaches to named entity recognition (NER) are based on sequential labeling methods, such as conditional random fields (CRFs), which label each word in a sentence and extract entities from them that correspond to named entities. With the extensive deployment of deep learning methods for sequential labeling tasks, state-of-the-art NER performance has been achieved on long short-term memory (LSTM) architectures using only basic features. In this paper, we address Korean NER tasks and propose an extension of a bidirectional LSTM CRF by investigating character-based representation. Our extension involves deploying a hybrid representation using ConvNet and LSTM for the sequential modeling of characters, namely a *character-based LSTM-ConvNet hybrid representation*. Using morphemes as processing units for bidirectional LSTM, we apply a proposed hybrid representation composed of morpheme vectors. Experimental results showed that the proposed LSTM-ConvNet hybrid representation yielded improvements over each single representation on standard Korean NER tasks.
Q3

Keywords: Named entity recognition; Long short term memory; Convolutional neural networks; Character-based composition

1. Introduction

Named entity recognition (NER) is an important subtask in information extraction (IE) (Sarawagi, 2008). It involves the extraction or identification of named entities in a domain world, such as person, organization, or location, or domain-specific entities such as diseases, dishes, or restaurants. When machine learning methods became popular for solutions to problems in artificial intelligence, commonly undertaken approaches in NER tasks were based on *sequential labeling* methods such as conditional random fields (CRFs) and structured SVM (support vector machine) in a supervised or semi-supervised way. In these approaches, named entities in text that need to be extracted are annotated with pre-defined labels and sequential labeling models are trained from annotated corpora. Given a test sentence, a sequential labeler from the trained model is applied to a sequence of words in the sentence, and the named entities are extracted by decoding the resulting labeled words. These sequential labeling methods

This paper has been recommended for acceptance by Pascale Fung
* Corresponding author.
E-mail address: nash@jbnu.ac.kr (S.-H. Na).

https://doi.org/10.1016/j.csl.2018.09.005
0885-2308/ 2018 Published by Elsevier Ltd.

have recently delivered state-of-the-art performance on the NER tasks (Hammerton, 2003; McCallum & Li, 2003; Lin & Wu, 2009; Collobert et al., 2011; Passos et al., 2014; Huang et al., 2015; Ma & Hovy, 2016; Chiu & Nichols, 2016; Lample et al., 2016).

While classical machine learning (ML) approaches are based largely on handcrafted features, deep learning reduces the burden on these feature engineering steps through a key capability called representation learning (Bengio, 2009). Representation learning enables a model to learn to compose complicated features from simpler ones. Successful methods in NER are based on feed-forward neural networks (Collobert et al., 2011), and long short-term memory (LSTM) (Huang et al., 2015; Ma & Hovy, 2016; Chiu & Nichols, 2016) and stack LSTM (Lample et al., 2016).

However, one of the limitations of applying deep learning to NER is the out-of-vocabulary (OOV) problem. Since the vectors representing word embeddings can be defined based on words observed in a corpus, such vectors are not available for unknown words that are relatively common in NER tasks. Specifically, the OOV problem becomes much more severe for morphologically rich languages such as Korean, because vocabulary sets are not bound; they are nearly unlimited.

To handle the OOV problem, compositional methods have been recently developed by deriving the vector of a word from character vectors of its compositional characters in a hierarchical manner (dos Santos & Zadrozny, 2014; Ling et al., 2015). One type of composition is based on convolutional neural networks (ConvNets) (dos Santos & Zadrozny, 2014) and another on bidirectional LSTMs (Ling et al., 2015). For NER tasks, Chiu and Nichols (2016); Ma & Hovy (2016) used ConvNet-based composition and Lample et al. (2016) deployed LSTM-based composition in a stack LSTM framework.

In this paper, we further explore compositional methods for Korean NER tasks by addressing the OOV problem. In our approach, we propose a novel *hybrid representation* that combines both LSTM-based and ConvNet-based compositional word vectors. To obtain this hybrid representation, we first separately apply LSTM and ConvNet-based compositions to input the character vectors and concatenate the resulting compositional morpheme vectors to finally generate the hybrid representation of a morpheme.

Experiment results on Korean NER tasks showed that our hybrid representation improved each of separate representations, implying that ConvNet- and LSTM-based compositions play different roles in capturing character-level features for NER tasks, thus making improvements in their combination.

The remainder of this paper is organized as follows: Section 2 describes related work in the area, and Section 3 details the proposed method as well as the corpus-processing method. Section 4 provides the experimental results, and our concluding remarks and a description of future work are provided in Section 5.

2. Related work

The commonly undertaken approach to NER involves regarding a task as a sequential labeling problem, called the *sequential labeling perspective*. From this perspective, a sequential tagger trained through supervised learning is applied to a sequence of words or characters in a given sentence to extract from it named entities of interest from among the labeled results. Since sequential labeling can also be seen as a machine learning method that addresses the problem of structured output, the NER literature on this perspective follows advances in machine learning. Previous studies on NER from this perspective are summarized in the following sections.

2.1. NER based on classical ML

Many studies are based on classical ML methods such as CRFs, the maximum entropy classifier, or SVM (Chieu & Ng, 2003; Klein et al., 2003; Zhang & Johnson, 2003; Ando & Zhang, 2005; Ratinov & Roth, 2009; Suzuki & Isozaki, 2008; Lin & Wu, 2009). The most important factors affecting performance in NER tasks have been discovered in the classical ML methods. For example, Chieu and Ng (2003) exploited discourse-level features and Klein et al. (2003) explored the effects of using character-level features.

Large-scale unlabeled texts have been extensively used to significantly improve performance in semi-supervised frameworks (Ando & Zhang, 2005; Suzuki & Isozaki, 2008) and through additional cluster-based features (Ratinov & Roth, 2009; Lin & Wu, 2009).

Please cite this article as: S. Na et al., Improving LSTM CRFs using character-based compositions for Korean named entity recognition, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.09.005