Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2017) xxx-xxx
www.elsevier.com/locate/csl

A Framework for pre-training hidden-unit conditional random fields and its extension to long short term memory networks ☆

Young-Bum Kim[a,*], Karl Stratos[b], Ruhi Sarikaya[a]

[a] *The Alexa Brain, Amazon, Seattle, WA, 98121, USA*
[b] *Toyota Technological Institute at Chicago, IL, USA*

Received 18 April 2016; received in revised form 24 December 2016; accepted 12 May 2017

Abstract

In this paper, we introduce a simple unsupervised framework for *pre-training* hidden-unit conditional random fields (HUCRFs), i.e., learning initial parameter estimates for HUCRFs prior to supervised training. Pre-training is generally important for models with non-convex training objectives such as deep neural nets. Our framework exploits the model structure of HUCRFs to make effective use of unlabeled data. The key idea is to use the separation of HUCRF parameters between observations and labels: this allows us to pre-train observation parameters independently of label parameters on resources such as unlabeled data or data labeled for non-target tasks. Pre-training is achieved by creating pseudo-labels from such resources. In the case of unlabeled data, we cluster observations and use the resulting clusters as pseudo-labels. Observation parameters can be trained on these external resources and then transferred to initialize the supervised training process on the target labeled data. Experiments on various sequence labeling tasks demonstrate that the proposed pre-training method consistently yields significant improvement in performance. The core idea could be extended to other learning techniques including deep learning. We applied the proposed technique to recurrent neural networks (RNN) with long short term memory (LSTM) architecture and obtained similar gains.

Keywords: Pre-training; Transfer learning; Spoken language understanding; Sequence labeling; Conditional random fiends; Multi-sense clustering; Word embedding; Hidden unit conditional random fields; LSTMs

1. Introduction

There has been a tremendous surge in the applications of machine learning and artificial intelligence (AI) to a number of problems during the past decade. In particular, cloud-driven personal digital assistants and AI systems are surfacing on many devices, including smart-phones, PCs, game consoles, headless speakers and wearables Sarikaya (2017). Information extraction, entity and slot tagging are essential tasks for numerous scenarios supported on these systems, where data from back-end knowledge bases and web services are presented to the user based on the semantic parsing of the natural language query. These tasks are considered as sequence labeling problems.

☆ This paper has been recommended for acceptance by Roger K. Moore.
* Corresponding author.
E-mail addresses: youngbum@amazon.com, stylebbum@gmail.com (Y.-B. Kim).

http://dx.doi.org/10.1016/j.csl.2017.05.004
0885-2308/ 2017 Published by Elsevier Ltd.

Over the past 15 years, conditional random fields (CRFs) (Lafferty et al., 2001; Sutton and McCallum, 2007) have been widely used for numerous sequence labeling problems (Collins, 2002; McCallum and Li, 2003; Sha and Pereira, 2003; Turian et al., 2010; Kim and Snyder, 2012; Celikyilmaz et al., 2013; Sarikaya et al., 2014; Anastasakos et al., 2014; Kim et al., 2014; 2015a; 2015d; 2015b; 2016d; 2016a; Yang et al., 2016). CRFs are appealing because of two reasons. First, the training objective considers entire sequences and can incorporate arbitrary features; furthermore, it is convex and can be optimized relatively efficiently using dynamic programming. Second, CRF performance is fairly task independent; it does not need significant tuning in model configurations and hyper-parameters to achieve competitive results.

Recently, deep learning techniques such as recurrent neural networks (RNNs) and its specific configuration long short term memory (LSTM) networks Hochreiter and Schmidhuber (1997) improved the state-of-the-art performance on many natural language processing and sequence modeling tasks (Kim et al., 2016c; Zhang and Wang, 2016; Guo et al., 2014; Liu and Lane, 2016; Chen et al., 2016; Kim et al., 2017a; Shi et al., 2015; Hori et al., 2016; Kim et al., 2017b; 2016b). Pre-training is an important step in training deep learning methods, as it helps model parameter estimation and leads to better model accuracy, particularly when labeled training data is limited Erhan et al. (2010).

So far, CRF family of learning methods lacked a framework to leverage unlabeled data to improve model parameter estimation and obtain improved accuracy. In this paper, we present a framework for pre-training the hidden unit CRFs (HUCRFs). The framework aims to learn the parameters of the HUCRFs in such a way to capture the deep structures and regularities either in the unlabeled data or through a related task. There could be many techniques to achieve this goal. We propose a technique that leverages a collection of unlabeled text, where the words are clustered into a set of clusters and cluster IDs are used as pseudo-labels to train HUCRFs. Then, the learned parameters corresponding to observations are used to initialize the training process on the labeled data. This pre-training step significantly reduces the challenges in training a highly accurate HUCRF by acquiring a broad feature coverage and finding a good initialization point. Our analysis implies that it also captures deep syntactic and semantic dependencies in the unlabeled data.

Additionally, we introduce a novel word clustering scheme based on canonical correlation analysis (CCA) that is sensitive to multiple word senses. For example, the resulting clusters will differentiate the occurrences of "bank" between financial institutions and the land alongside water. This is an important point as different senses of a word are likely to have a different task specific tag. Putting them in different clusters would enable the HUCRF model to learn the distinction in terms of label assignment.

This work was originally published in Kim et al. (2015c; 2015d). The main contributions in this paper are:

- We combined previous publication with the detailed explanation and in-depth analysis of the results.
- We added experiments with public data set for reproducibility of our results and provide direct comparison to the previous research.
- We also applied the proposed pre-training technique to deep learning, in particular, to recurrent neural networks (RNNs) with long short term memory (LSTM) architecture and achieved similar improvements.

2. Related work

Even though CRFs have been widely used on numerous sequence learning tasks, they still lack in certain aspects. For example, the linearity of CRFs is limiting: their expressive power is limited by the inner product between data and model parameters. Although one can still alleviate this limitation, for example by using non-linear data representations as features, it may be desirable to consider a richer sequence labeling paradigm. Several previous studies (Quattoni et al., 2007; Maaten et al., 2011; Stratos et al., 2013) proposed non-linear models by adding latent variables to the sequence modeling. However, increased expressive power typically results in intractable training objectives, and approximations are often used. In practice, non-linear models are very successful, as testified by many recent empirical breakthroughs with deep learning methods particularly LSTMs Lample et al. (2016).

In the non-linear models, it is often very useful to make use of unlabeled data for obtaining initial parameter values Particularly in limited labeled data scenarios, the best models across different tasks mostly exploit an unsupervised pre-training strategy followed by the supervised training phase. This model training recipe is considered as a

Please cite this article as: Y. Kim et al., A Framework for pre-training hidden-unit conditional random fields and its extension to long short term memory networks, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.05.004