ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language xxx (2014) xxx–xxx

COMPUTER
SPEECH AND
LANGUAGE

www.elsevier.com/locate/csl

Recurrent neural network language model adaptation with curriculum learning☆

Q1 Yangyang Shi*, Martha Larson, Catholijn M. Jonker

Q2 *Department of Intelligent Systems, Delft University of Technology, Mekelweg 4, 2628CD Delft, The Netherlands*

Received 11 September 2013; received in revised form 13 September 2014; accepted 13 November 2014

Abstract

This paper addresses the issue of language model adaptation for recurrent neural network language models (RNNLMs), which have recently emerged as a state-of-the-art method for language modeling in the area of speech recognition. Curriculum learning is an established machine learning approach that achieves better models by applying a curriculum, i.e., a well-planned ordering of the training data, during the learning process. Our contribution is to demonstrate the importance of curriculum learning methods for adapting RNNLMs and to provide key insights on how it should be applied. RNNLMs model language in a continuous space and can theoretically exploit word-dependency information over arbitrarily long distances. These characteristics give RNNLMs the ability to learn patterns robustly with relatively little training data, implying that they are well suited for adaptation. In this paper, we focus on two related challenges facing language models: *within-domain adaptation* and *limited-data within-domain adaptation*. We propose three types of curricula that start with general data, i.e., characterizing the domain as a whole, and move towards specific data, i.e., characterizing the sub-domain targeted for adaptation. Effectively, these curricula result in a model that can be considered to represent an implicit interpolation between general data and sub-domain-specific data. We carry out an extensive set of experiments that investigates how adapting RNNLMs using curriculum learning can improve their performance.

Our first set of experiments addresses the within-domain adaptation challenge, i.e., creating models that are adapted to specific sub-domains that are part of a larger, heterogeneous domain of speech data. Under this challenge, all training data is available to the system at the time when the language model is trained. First, we demonstrate that curriculum learning can be used to create effective sub-domain-adapted RNNLMs. Second, we show that a combination of sub-domain-adapted RNNLMs can be used if the sub-domain of the target data is unknown at test time. Third, we explore the potential of applying combinations of sub-domain-adapted RNNLMs to data for which sub-domain information is unknown at training time and must be inferred.

Our second set of experiments addresses limited-data within-domain adaptation, i.e., adapting an existing model trained on a large set of data using a smaller amount of data from the target sub-domain. Under this challenge, data from the target sub-domain is not available at the time when the language model is trained, but rather becomes available little by little over time. We demonstrate that the implicit interpolation carried out by applying curriculum learning methods to RNNLMs outperforms conventional interpolation and has the potential to make more of less adaptation data.

Keywords: Recurrent neural networks; Language models; Curriculum learning; Latent Dirichlet allocation; Topics; Socio-situational setting

☆ This paper has been recommended for acceptance by J. Glass.

Q3 * Corresponding author. Tel.: +31 0681861586.
E-mail address: yangyangshi@ieee.org (Y. Shi).

http://dx.doi.org/10.1016/j.csl.2014.11.004
0885-2308/© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

The task of statistical language models is to judge whether a sequence of words is well formed or not. Conventional n-gram language models factorize the joint probabilities of all the words in a sequence into a product of probabilities of each word given information about its history, i.e., the preceding $n-1$ words. By using word histories, n-gram language models capture local regularities of languages. However, n-gram language models can only exploit an n-gram if the exact string of n words is present in the training data. As n grows large, the chance that an n-gram seen in the target data was also present in the training data falls off sharply. For this reason, conventional n-gram language models easily suffer from data sparseness. In practice, the history length $n-1$ that can be effectively exploited is quite limited. For this reason, n-gram language models lack adequate means to model long-distance dependencies.

These known shortcomings are addressed by recurrent neural network language models (RNNLMs). Recently, RNNLMs have been demonstrated to outperform n-gram language models for speech recognition (Mikolov et al., 2010). Their superior capabilities rely on two mechanisms. First, RNNLMs map the discrete word-based vocabulary into a continuous space. As a result, the model can exploit word sequences which are similar, without requiring them to be exactly identical. This mechanism helps to reduce the effect of data sparseness. Second, RNNLMs are explicitly equipped to handle long-distance dependencies. The recurrent loop in their architecture constitutes a memory that allows them to model arbitrarily long word histories theoretically.

In this paper, we investigate language model adaptation for RNNLMs, and specifically address two central challenges for language model adaptation, *within-domain adaptation* and *limited-data within-domain adaptation*, originally identified by Rosenfeld (1994) and explained later in depth. The main contribution of this paper is to demonstrate that curriculum learning is an important technique for carrying out the adaptation of RNNLMs and to provide insights on how it must be applied in order to be effective for improving speech recognition.

Curriculum learning applies a specific, well-planned ordering of the training data, referred to as a 'curriculum', during the learning process and is an established approach in the machine learning community. When conventional n-gram language models are trained, the order in which the training data is processed has no impact on the outcome of the training process. In contrast, neural networks are indeed sensitive to the differences in the order in which the training data is presented to them. The work of Bengio et al. (2009) attributes the benefits of curriculum learning in neural network training to an ability of the curriculum to guide the learner, in particular, directing it away from inappropriate local minima and towards more suitable ones.

The advantages that curriculum learning offers to RNNLMs for speech recognition have been previously established in the literature (Mikolov et al., 2010, 2011). The previous work has focused on dynamically updating language models during the recognition process (Mikolov et al., 2010) and in optimal reduction and sorting of the training data (Mikolov et al., 2011). The existing work points out that training data presented later in the training process has more influence on the final form of the model than the initial part of the training data. As such, curriculum learning can be used to accomplish an implicit interpolation of the training data, where certain parts of the data are given more importance than others.

The specific issue of adaptation is particularly important for RNNLMs, and as such deserves dedicated attention. A key reason for its importance is the relatively high cost of training RNNLMs, which can be attributed to a range of factors. Here, we mention in particular the fact that the whole training set is usually presented to the model multiple times (referred to as 'training epochs'). The relatively high cost of the training phase of RNNLMs means that retraining the language model whenever new training data becomes available is prohibitively costly.

Curriculum learning has several distinction advantages to offer for the adaptation of RNNLM to specific subdomains. First, curriculum learning provides a method to effectively carry out implicit interpolation that does not require the parameter optimization needed for conventional interpolation. Second, as has been pointed out by Iyer et al. (1994), Iyer and Ostendorf (1999), and, more recently, by Mikolov and Zweig (2012), one danger of adaptation models that build individual component models on data sub-sets is fragmentation. Fragmentation refers to the fact that as more and more component models are built, relatively less data is available to train each. Using curriculum learning to train sub-domain adapted RNNLMs neatly circumvents the fragmentation issue. All training data can be used to train each adapted model; it is the order in which the data is presented to the model during training that makes the difference. In short, although the potential of curriculum learning for RNNLMs has been established and offers clear advantages, the

Please cite this article in press as: Shi, Y., et al., Recurrent neural network language model adaptation with curriculum learning. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.11.004