Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 40 (2016) 1–22

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Semantic language models with deep neural networks

Ali Orkan Bayer *, Giuseppe Riccardi

Signals and Interactive Systems Lab, Via Sommarive, 5 38123 Povo, Trento, Italy

Received 22 October 2015; received in revised form 29 February 2016; accepted 17 April 2016
Available online 4 May 2016

Abstract

In this paper we explore the use of semantics in training language models for automatic speech recognition and spoken language understanding. Traditional language models (LMs) do not consider the semantic constraints and train models based on fixed-sized word histories. The theory of frame semantics analyzes word meanings and their constructs by using "semantic frames". Semantic frames represent a linguistic scene with its relevant participants and their relations. They are triggered by target words and include slots which are filled by frame elements. We present semantic LMs (SELMs), which use recurrent neural network architectures and the linguistic scene of frame semantics as context. SELMs incorporate semantic features which are extracted from semantic frames and target words. In this way, long-range and "latent" dependencies, i.e. the implicit semantic dependencies between words, are incorporated into LMs. This is crucial especially when the main aim of spoken language systems is understanding what the user means. Semantic features consist of low-level features, where frame and target information is directly used; and deep semantic encodings, where deep autoencoders are used to extract semantic features. We evaluate the performance of SELMs on publicly available corpora: the Wall Street Journal read-speech corpus and the LUNA human–human conversational corpus. The encoding of semantic frames into SELMs improves the word recognition performance and especially the recognition performance of the target words, the meaning bearing elements of semantic frames. We assess the performance of SELMs for the understanding tasks and we show that SELMs yield better semantic frame identification performance compared to recurrent neural network LMs.

Keywords: Language modeling; Recurrent neural networks; Frame semantics; Semantic language models; Deep autoencoders

1. Introduction

Statistical language models (LMs) are one of the main knowledge sources in language processing systems such as statistical machine translation, information retrieval, automatic speech recognition (ASR), and spoken language understanding (SLU). They play a crucial role in searching for the best hypothesis by estimating the likelihood of each hypothesis in that language.

Traditional LMs are based on n-gram models, where the probability of a word is only dependent on the previous $(n-1)$ words. Although currently the state-of-the-art performance for LMs is obtained by using recurrent neural networks (RNNs), n-gram LMs are still widely used because of their simple computational architecture and also because they scale well (Chelba et al., 2012, 2013) on large data. However, since n-gram LMs are trained by considering fixed

* Corresponding author at: Signals and Interactive Systems Lab, Via Sommarive, 5 38123 Povo, Trento, Italy. Tel.: +39 3209714240; fax: +39 0461283166.

E-mail address: aliorkan.bayer@unitn.it (A.O. Bayer), giuseppe.riccardi@unitn.it (G. Riccardi).

http://dx.doi.org/10.1016/j.csl.2016.04.001

size histories of words, they suffer from the “locality problem” (Bellegarda, 2000a). As suggested by Bellegarda (2000a), this problem can be solved by “span extension”. Span extension can be performed either by using syntactic dependencies or semantic relations. This paper explores the use of lexical semantics for improving the performance of LMs.

One of the approaches that incorporates semantic information into LMs is the *topic model* (Gildea and Hofmann, 1999; Schwartz et al., 1997). Topics may be selected from a hand-crafted set or can be learned by data-driven approaches. Topic LMs model the probability of a word in a topic and do not consider the local structure of the language. Therefore they are combined with n-gram models.

Another approach for semantic span extension is to use trigger pairs. Trigger pairs capture semantic relations by using the correlation between words (Rosenfeld, 1996). Therefore if a word sequence A is significantly correlated with a word sequence B, they constitute a trigger pair. Then, A is referred to as the *trigger* and B as the *triggered sequence*. Rosenfeld (1996) reports that *self triggers* are very powerful and robust. Also trigger pairs of frequent words have more potential than the trigger pairs of infrequent words. Trigger pairs are determined by using the average mutual information between the trigger and the triggered sequence. Trigger pairs are very effective to handle the long-range dependencies. However, selection of trigger pairs is an issue.

Bellegarda (2000a, 2000b) uses *latent semantic analysis* (LSA) to extend the trigger pairs approach. LSA (Berry et al., 1995; Deerwester et al., 1990) is used as an indexing mechanism in information retrieval, and it maps the discrete space of words and documents to the same continuous space. Therefore, each word and each document is represented as a vector in this space. A word-document matrix, in which each column represents a document and each row represents a word, is constructed by populating the matrix values by normalized counts of the words in the documents. The normalization is done with respect to the number of documents in which the word occurs. Because of the computational requirements, singular value decomposition is applied to this matrix. The final representation conceptually represents each word and document as a linear combination of abstract concepts, which is very similar to the distributed representations the neural network LMs use. At the final step, LMs are modeled over the LSA history of the word. Combining LSA with n-gram models resulted in significant improvements in perplexity and word error rate (Bellegarda, 2000a, 2000b).

Traditional LMs also suffer from “curse of dimensionality” (Bengio et al., 2003), i.e. they consider words as sequence of symbols and do not model the semantic relationships between these words. This problem is approached by learning distributed representations (also known as *word embeddings*) of words (Bengio et al., 2003), i.e. words are mapped onto a continuous space. Neural network LMs (NNLMs) are first introduced in Bengio et al. (2003). NNLMs learn and use distributed representations of words in language modeling. NNLMs are reported to reduce the perplexity significantly (Bengio et al., 2003). Also Schwenk (2007) applies NNLMs to ASR which reports significant improvements in word error rate (WER) by linearly interpolating NNLMs with back-off n-gram LMs. The first approach to NNLMs were feed-forward NNLMs which are based on fixed size histories; therefore they also suffer from the problems related to fixed size histories. Recurrent NNLMs (RNNLMs) (Mikolov, 2012; Mikolov et al., 2010), overcome this problem by using recurrent connections, which feed the activation of the hidden layer at the previous time step as input. This can be thought of as a short-term memory which enables the network to model long-range histories. RNNLMs are shown to improve perplexities and WERs better than any feed-forward NNLM (Mikolov et al., 2011a). Mikolov and Zweig (2012) add a feature layer to RNNLMs, where topic features are used as additional context to the NNLM. In addition, Mikolov et al. (2011b) present the training of maximum entropy features jointly with the RNNLM which are referred to as RNNME.

When building LMs for a specific application, LMs are tuned with respect to the performance metric of the target application. This may lead to problems especially for spoken dialog systems, where one of the main goals of these systems is to extract user intentions and the meaning of utterances. Spoken dialog systems most often use a cascaded approach, where the output of the ASR is fed into the SLU module. The LMs that are used in ASR and SLU are optimized with respect to the component that they are trained for. LMs for ASR are optimized to lower the WER. LMs are optimized to lower the concept error rate (CER) in SLU. In the literature, it has been argued that LMs should be optimized jointly, since the best recognition performance does not yield the best understanding performance (Deoras et al., 2013; Riccardi and Gorin, 1998; Wang et al., 2003). Therefore, it is important that LMs are trained by considering semantic constraints in that language.

As we have mentioned traditional n-gram LMs suffer from the “locality problem” and “curse of dimensionality”. The semantic span extension approaches like the topic model (Gildea and Hofmann, 1999; Schwartz et al., 1997) can only be used through a combination of the model with an n-gram LMs. Trigger pairs (Rosenfeld, 1996) offer an