Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 29 (2015) 20–31

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Unsupervised language model adaptation using LDA-based mixture models and latent semantic marginals☆

Md. Akmal Haidar*, Douglas O'Shaughnessy

INRS-EMT, 800 de la Gauchetiere Ouest, Bureau 6900, H5A 1K6 Montreal, QC, Canada

Received 17 December 2013; received in revised form 16 June 2014; accepted 18 June 2014
Available online 2 July 2014

Abstract

In this paper, we present unsupervised language model (LM) adaptation approaches using latent Dirichlet allocation (LDA) and latent semantic marginals (LSM). The LSM is the unigram probability distribution over words that are calculated using LDA-adapted unigram models. The LDA model is used to extract topic information from a training corpus in an unsupervised manner. The LDA model yields a document–topic matrix that describes the number of words assigned to topics for the documents. A hard-clustering method is applied on the document–topic matrix of the LDA model to form topics. An adapted model is created by using a weighted combination of the *n*-gram topic models. The stand-alone adapted model outperforms the background model. The interpolation of the background model and the adapted model gives further improvement. We modify the above models using the LSM. The LSM is used to form a new adapted model by using the minimum discriminant information (MDI) adaptation approach called unigram scaling, which minimizes the distance between the new adapted model and the other model. The unigram scaling of the adapted model using LSM yields better results over a conventional unigram scaling approach. The unigram scaling of the interpolation of the background and the adapted model using the LSM outperform the background model, the unigram scaling of the background model, the unigram scaling of the adapted model, and the interpolation of the background and the adapted models respectively. We perform experiments using the '87–89 Wall Street Journal (WSJ) corpus incorporating a multi-pass continuous speech recognition (CSR) system. In the first pass, we used the background *n*-gram language model for lattice generation and then we apply the LM adaptation approaches for lattice rescoring in the second pass.

Keywords: Language model; Topic model; Mixture model; Speech recognition; Minimum discriminant information

1. Introduction

LM adaptation plays a vital role to improve a speech recognition system's performance. It is essential when the styles, topics or domains of the recognition tasks are mismatched with the training set. To compensate for this mismatch, LM adaptation helps to exploit specific, albeit limited, knowledge about the recognition task (Bellegarda, 2004). The idea of an unsupervised LM adaptation approach is to extract latent topics from the training set and then adapt

☆ This paper has been recommended for acceptance by R. De Mori.

* Corresponding author. Tel.: +1 5149951266.
E-mail addresses: haidar@emt.inrs.ca (Md.A. Haidar), dougo@emt.inrs.ca (D. O'Shaughnessy).

http://dx.doi.org/10.1016/j.csl.2014.06.002
0885-2308/

topic-specific LMs with proper mixture weights, finally interpolated with a generic *n*-gram LM (Liu and Liu, 2007; Haidar and O'Shaughnessy, 2010).

Statistical *n*-gram language models have been used successfully for speech recognition and other applications. They use local context information by modeling text as a Markovian sequence and capture only the local dependencies between words. They suffer from insufficiencies of the training data, which limit model generalization. Due to limitations of the amount of training data, statistical *n*-gram LMs encounter a data sparseness problem, which is handled by using backoff smoothing approaches with lower-order language models (Chen and Goodman, 1999). Moreover, *n*-gram models cannot capture the long-range information of natural language. Several methods have been investigated to overcome this weakness. A cache-based language model is an earlier approach that is based on the idea that if a word appeared previously in a document it is more likely to occur again. It helps to increase the probability of previously observed words in a document when predicting a future word (Kuhn and Mori, 1990). Recently, various techniques such as latent semantic analysis (LSA) (Deerwester et al., 1990; Bellegarda, 2000), probabilistic LSA (PLSA) (Gildea and Hofmann, 1999), and LDA (Blei et al., 2003) have been investigated to extract the latent topic information from a training corpus. All of these methods are based on a bag-of-words assumption, i.e., the word-order in a document can be ignored. In LSA, a word–document matrix is used to extract the semantic information. In PLSA, each document is modeled by its own mixture weights and there is no generative model for these weights. So, the number of parameters grows linearly when increasing the number of documents, which leads to an overfitting problem. Also, there is no method to assign probability for a document outside the training set. On the contrary, the LDA model was introduced where a Dirichlet distribution is applied on the topic mixture weights corresponding to the documents in the corpus. Therefore, the number of model parameters is dependent only on the number of topic mixtures and the vocabulary size. Thus, LDA is less prone to overfitting and can be used to compute the probabilities of unobserved test documents. However, the LDA model can be viewed as a set of unigram latent topic models. The LDA model has been used successfully in recent research work for LM adaptation (Tam and Schultz, 2005, 2006; Liu and Liu, 2007, 2008; Haidar and O'Shaughnessy, 2010, 2011, 2012b,a). In Tam and Schultz (2006), a unigram scaling approach is used for the LDA adapted unigram model to minimize the distance between the adapted model and the background model (Tam and Schultz, 2006). The LDA model is also used as a clustering algorithm to cluster training data into topics (Ramabhadran et al., 2007; Heidel and Lee, 2007). The LDA model can be merged with *n*-gram models and achieve perplexity reduction (Sethy and Ramabhadran, 2008). A non-stationary version of LDA can be developed for LM adaptation in speech recognition (Chueh and Chien, 2009). A topic-dependent LM, called topic dependent class (TDC) based *n*-gram LM, was proposed in Naptali et al. (2012), where the topic is decided in an unsupervised manner. Here, the LSA method was used to reveal latent topic information from noun–noun relations (Naptali et al., 2012).

The simple technique to form a topic from an unlabeled corpus is to assign one topic label to a document (Iyer and Ostendorf, 1996). This hard-clustering strategy is used with leveraging LDA and named entity information to form topics (Liu and Liu, 2007, 2008). Here, topic-specific *n*-gram language models are created and joined with proper mixture weights for adaptation. The adapted model is then interpolated with the background model to capture the local lexical regularities. The component weights of the *n*-gram topic models were created by using the word counts of the latent topic of the LDA model. However, these counts are best suited for the LDA unigram topic models. A unigram count weighting approach (Haidar and O'Shaughnessy, 2010) for the topics generated by hard-clustering has shown better performance over the weighting approach described in Liu and Liu (2007, 2008). An extension of the unigram weighting approach (Haidar and O'Shaughnessy, 2010) was proposed in Haidar and O'Shaughnessy (2011) where the weights of the *n*-gram topic models are computed by using the *n*-gram count of the topics generated by a hard-clustering method. The adapted *n*-gram model is scaled by using the LDA-adapted unigram model called latent semantic marginals (LSM) (Tam and Schultz, 2006) and outperforms a traditional unigram scaling of the background model using the above marginals (Haidar and O'Shaughnessy, 2012a). Here, the unigram scaling technique (Kneser et al., 1997) is applied where a new adapted model is formed by using a minimum discriminant information (MDI) approach that minimizes the KL divergence between the new adapted model and the adapted *n*-gram model, subject to a constraint that the marginalized unigram distribution of the new adapted model is equal to the LSM. In this paper, we present an extension to the previous works (Haidar and O'Shaughnessy, 2011, 2012a) where we apply the unigram scaling technique to the interpolation of the background and the adapted *n*-gram model and note better results over the previous works. In addition, we perform all the experiments using different corpus sizes ('87 WSJ corpus (17 million words) and '87–89 WSJ corpus (37 million words)) instead of using only the 1 million words WSJ training