Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 192–209

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Semantic spaces for improving language modeling☆

Tomáš Brychcín[a,b,*], Miloslav Konopík[a,b]

[a] *Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*

[b] *NTIS – New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*

Received 21 March 2012; received in revised form 29 March 2013; accepted 10 May 2013
Available online 19 May 2013

Abstract

Language models are crucial for many tasks in NLP (Natural Language Processing) and n-grams are the best way to build them. Huge effort is being invested in improving n-gram language models. By introducing external information (morphology, syntax, partitioning into documents, etc.) into the models a significant improvement can be achieved. The models can however be improved with no external information and smoothing is an excellent example of such an improvement.

In this article we show another way of improving the models that also requires no external information. We examine patterns that can be found in large corpora by building semantic spaces (HAL, COALS, BEAGLE and others described in this article). These semantic spaces have never been tested in language modeling before. Our method uses semantic spaces and clustering to build classes for a class-based language model. The class-based model is then coupled with a standard n-gram model to create a very effective language model.

Our experiments show that our models reduce the perplexity and improve the accuracy of n-gram language models with no external information added. Training of our models is fully unsupervised. Our models are very effective for inflectional languages, which are particularly hard to model. We show results for five different semantic spaces with different settings and different number of classes. The perplexity tests are accompanied with machine translation tests that prove the ability of proposed models to improve performance of a real-world application.

Keywords: Class-based language models; Semantic spaces; HAL; COALS; BEAGLE; Random Indexing; Purandare and Pedersen; Clustering; Inflectional languages; Machine translation

1. Introduction

Language modeling is a crucial task in many areas of NLP. Speech recognition, optical character recognition and many other areas heavily depend on the performance of the language model that is being used. Each improvement in language modeling may also improve the particular job where the language model is used.

☆ This paper has been recommended for acceptance by 'Dr. E. Briscoe'.

* Corresponding author at: Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic. Tel.: +420 377632418.

E-mail addresses: brychcin@kiv.zcu.cz (T. Brychcín), konopik@kiv.zcu.cz (M. Konopík).

0885-2308/$ – see front matter
http://dx.doi.org/10.1016/j.csl.2013.05.001

Research into language modeling started more than 20 years ago and has evolved into a very mature discipline. Now it is very difficult to outperform the state of the art. Our research is focused on inflectional languages as we believe that these languages offer some room for improvement. We however also provide experiments for English (which is not a very inflectional language). Even in the case of English, we were able to obtain positive results.

Czech and Slovak belong to the Slavic language group. These languages are highly inflectional and have a relatively free word order. Czech has seven cases and three genders. Slovak has six cases and also three genders. The word order is very variable from the syntactic point of view: words in a sentence can usually be ordered in several ways, each carrying a slightly different meaning. These properties of the languages complicate the language modeling task. The great number of word forms and more possible word sequences lead to a greater number of n-grams. Data sparsity is a common problem of language models. In Czech, Slovak and other Slavic languages, this problem is more evident.

Class-based modeling is the most popular technique used for reducing the huge vocabulary-related sparseness of statistical language models (Brown et al., 1992). Individual words are clustered into a much smaller number of classes. As a result, less data are required to train a robust class-based language model. Both manual and automatic word-clustering techniques are being used. Standalone class-based models usually perform poorly, which is the reason why they are usually combined with other models. Many researchers have demonstrated that the combination of a standalone class-based language model and a standard word n-gram model reduces the model perplexity (Maltese et al., 2001; Whittaker, 2000; Whittaker and Woodland, 2003).

An effective solution for language modeling is to use information about the morphology of the language. In Oparin (2008) experiments with morphological random forests in the Czech and Russian language are shown with the conclusion that they can be used effectively for inflectional languages. Authors of Vaiciunas et al. (2004) describe the language modeling of Lithuanian by means of class-based language models derived by word clustering and morphological word decomposition and their linear interpolation with the baseline word n-gram model. The authors present a perplexity reduction of 8–13% depending on the size of the corpora. A similarly effective solution is to use class-based language models where classes are derived from lemmas and morphological categories (Brychcín and Konopík, 2011). The article shows a perplexity reduction of 10–30% in corpora in the Czech and Slovak languages. A comparative study of several methods using morphological information for modeling conversational Arabic can be found in Kirchhoff et al. (2006). The usage of morphological information seems to be very effective for inflectional languages; however, it requires a huge number of manually annotated texts.

In Brown et al. (1992) the MMI (Maximum Mutual Information) clustering algorithm was introduced. This algorithm is based upon the principle of merging a pair of words into one class according to the minimal mutual information loss principle. The algorithm gives very satisfactory results and it is completely unsupervised. Its complexity is however very problematic. This method of word clustering is possible only in very small corpora and is not suitable for large vocabulary applications. The authors in Yokoyama et al. (2003) used the MMI algorithm to build class-based language models. Their linear interpolation with the word n-gram model was applied to speech recognition of Japanese. The authors showed a 2% absolute improvement in word accuracy but only in very small corpora.

Several authors have tried to approximate the MMI algorithm to reduce computational requirements and to make it more suitable for large vocabulary language models (Bai et al., 1998; Yamamoto and Sagisaka, 1999). Automatically derived clusters have been used for class-based language models of Japanese and Chinese (Gao et al., 2002). The authors concentrated on the best way of using the clusters; however, they did not focus on how to get them.

Another way of improving language models is to use semantic information. This idea is based on the assumption that words with lexically different forms usually share similar meanings in cases where they frequently occur in similar contexts. The semantic information can be calculated using the Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Landauer and Dumais, 1997; Landauer et al., 1998) method or its probabilistic variant, the PLSA (Hofmann, 1999). A similar method to the PLSA is the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) which is essentially the Bayesian version of the PLSA model. Bellegarda and his team were the first to introduce LSA into language modeling (Bellegarda et al., 1996). Their approach consisted in using LSA to derive word clusters for class-based language models.

The approach then evolved to focus on documents instead of focusing on words. It is assumed that documents may vary in domain, topic and styles, which means that they also differ in the probability distribution of n-grams. This assumption is used for adapting language models to the long context (domain, topic, style of particular documents). LSA (or similar methods) are used to find out which documents are similar and which are not. This long context information is added to standard n-gram models to improve their performance. A very effective group of models (sometimes