Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 33 (2015) 88–108

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Latent semantics in language models☆

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 5 December 2013; received in revised form 2 January 2015; accepted 15 January 2015
Available online 22 January 2015

Abstract

This paper investigates three different sources of information and their integration into language modelling. Global semantics is modelled by Latent Dirichlet allocation and brings long range dependencies into language models. Word clusters given by semantic spaces enrich these language models with short range semantics. Finally, our own stemming algorithm is used to further enhance the performance of language modelling for inflectional languages.

Our research shows that these three sources of information enrich each other and their combination dramatically improves language modelling. All investigated models are acquired in a fully unsupervised manner.

We show the efficiency of our methods for several languages such as Czech, Slovenian, Slovak, Polish, Hungarian, and English, proving their multilingualism. The perplexity tests are accompanied by machine translation tests that prove the ability of the proposed models to improve the performance of a real-world application.

Keywords: Language models; Latent Dirichlet allocation; Semantic spaces; Stemming; HAL; COALS; Random indexing; HPS; LDA; Machine translation; Moses

1. Introduction

Language modelling is an essential part in many tasks of natural language processing (NLP). Speech recognition, machine translation, optical character recognition, and many other disciplines strongly depend on the language model and thus every improvement in language modelling can also improve the performance of the whole system.

In this paper we explore fully unsupervised methods for language modelling (which require no labelled data and no information about language itself). To prove their multilingualism we experiment with several languages including highly inflectional as well as low-inflection languages. We incorporate three different families of languages (Slavic, Uralic, and Germanic) into our experiments. As representatives of Slavic languages we experiment with Czech,

☆ This paper has been recommended for acceptance by 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).

http://dx.doi.org/10.1016/j.csl.2015.01.004
0885-2308/

hockey, Pittsburgh, Jágr, goal, 68	hockey, football, soccer, basketball, tennis
win, champion, medal, gold, best	win, lose, play, wager, defeat
(a) Global context semantics	(b) Local context semantics

Fig. 1. Examples of semantically similar words. Each row represents semantically similar words according to (a) global semantics with long range dependencies (words in the same line are likely to occur in similar contexts, but not at the same position) and (b) local semantics with short range dependencies (words in the same line should be mutually substitutable at the same position in the appropriate context).

Slovenian, Slovak, and Polish. Uralic languages are represented by Hungarian, and Germanic languages by English. All languages we investigate in this paper except English are characterized by a high level of inflection and relatively free word order (from the syntactic point of view, the words in a sentence can usually be reordered in several ways to carry a slightly different meaning). Properties of these languages complicate the language modelling task. The great number of word forms and large number of possible word sequences lead to a much higher number of n-grams. Data sparsity is a common problem of language models, but for highly inflected languages this problem is even more evident.

The highly inflected languages in this paper belong rather among non-mainstream languages, for which the language modelling task has not gained as much attention as it has for English, for example. We thus believe there is considerable potential for improvements. However we provide experiments also for English to compare our methods with the state of the art.

In this paper we extend our work on the application of semantic spaces in language modelling (Brychcín and Konopík, 2014), where we have achieved significant improvements in perplexity and in machine translation task especially with HAL, COALS and RI models. Thus, these models are investigated more deeply in this paper.

We attempt to improve language modelling by adding long-range semantic dependencies. We choose latent Dirichlet allocation (LDA) (Blei et al., 2003) for that task, because it has already been shown by many researchers that LDA improves language modelling (see, e.g., Tam and Schultz, 2005, 2006; Watanabe et al., 2011).

The performance of these language models is further enhanced by our unsupervised stemming algorithm called high precision stemmer (HPS)[1] introduced in Brychcín and Konopík (2015). We have already tested our stemmer in language modelling tasks and the results indicate that HPS performs best compared to other unsupervised stemmers.

To the best of our knowledge we are first to try to combine these three sources of information (i.e. local semantics, global semantics, and morphology).

2. State of the art in latent semantics

In the context of this article, we work with various methods for modelling semantic relations between words, and use them to improve our language models. The backbone principle of methods for discovering hidden meaning from a plain text is the formulation of distributional hypothesis in Firth (1957) that says *"a word is characterized by the company it keeps"*. The direct implication of this hypothesis is that the word meaning is related to the context where it usually occurs and thus it is possible to compare the meanings of two words by statistical comparisons of their contexts. This implication was confirmed by empirical tests carried out on human groups in Charles (2000).

Several authors have made huge efforts to give an overview of the current state of the art in computational methods for extracting meaning from text (Turney and Pantel, 2010; Riordan and Jones, 2011; McNamara, 2011).

All the methods for extracting meaning can be approximately summarized in two categories. Authors Riordan and Jones (2011) and McNamara (2011) categorize these methods into *context-word* and *context-region* approaches. In this paper we use the notation *local context* and *global context*, respectively, because we think this notation describes the principle of meaning extraction better. These two categories are briefly described in the following Sections 2.1 and 2.2. Additionally, to give a better idea of how these two approaches differ, Fig. 1 shows an example of global context and local context semantics of words.

Models based upon the distributional hypothesis usually represent the meaning as a point in a multi-dimensional space. Thus, one meaning is represented as a single vector. These models are then referred to as the vector-space models

[1] A description of the algorithm and its implementation is available at http://liks.fav.zcu.cz/HPS.