Acta Psychologica 133 (2010) 234–243

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Combining feature norms and text data with topic models

Mark Steyvers *

Department of Cognitive Sciences, 3151 Social Sciences Plaza, University of California, Irvine, CA 92697-5100, USA

ARTICLE INFO

Article history:
Received 2 January 2009
Received in revised form 20 October 2009
Accepted 29 October 2009
Available online 30 November 2009

PsycINFO classification:
4100
2340
2343

Keywords:
Semantic cognition
Semantic spaces
Feature representations
Feature norms
Topic models
Background knowledge
Bayesian models

ABSTRACT

Many psychological theories of semantic cognition assume that concepts are represented by features. The empirical procedures used to elicit features from humans rely on explicit human judgments which limit the scope of such representations. An alternative computational framework for semantic cognition that does not rely on explicit human judgment is based on the statistical analysis of large text collections. In the topic modeling approach, documents are represented as a mixture of learned topics where each topic is represented as a probability distribution over words. We propose feature-topic models, where each document is represented by a mixture of learned topics as well as predefined topics that are derived from feature norms. Results indicate that this model leads to systematic improvements in generalization tasks. We show that the learned topics in the model play in an important role in the generalization performance by including words that are not part of current feature norms.

1. Introduction

Featural representations have played a central role in psychological theories of semantic cognition and knowledge organization (Collins & Quillian, 1969; McRae, de Sa, & Seidenberg, 1997; Rogers & McClelland, 2004; Smith, Shoben, & Rips, 1974; Vigliocco, Vinson, Lewis, & Garrett, 2004). Many of these theories assume that the meaning of a concept can be represented by a set of features (also referred to as properties or attributes). Many behavioral experiments have been conducted to elicit detailed knowledge of features (e.g. De Deyne et al., 2008; McRae, Cree, Seidenberg, & McNorgan, 2005; Ruts et al., 2004; Vinson & Vigliocco, 2008). In a typical procedure, the subjects are asked to generate a list of features associated with a concept which might be followed by a verification stage in which the subject verifies which concepts are associated with a particular feature (e.g. De Deyne et al., 2008). Because the feature norming methods rely on explicit human judgment, it takes a large effort to build such databases. To date, feature norms have only been developed for a few hundred words. This limits the scope of any computational model for semantic cognition that is based on these feature norms. Also, it is not clear how people generate features in the generation task and whether all the features listed are relevant to understand mental representations (Zeigenfuse & Lee, 2008, 2010).

An alternative computational framework for semantic cognition that does not rely on explicit human judgment is based on the statistical analysis of large text collections. These models learn in an unsupervised fashion and require no external knowledge databases such as dictionaries, thesauri and other knowledge repositories. In this framework, information about the meaning of words can be derived by analyzing the co-occurrences between words and the contexts in which they occur (such as paragraphs or documents in a corpus of text). Many statistical text models for semantic cognition work with a "bag-of-words" representation, where each document is represented by vectors that contain the counts of the number of times each term (i.e., word or word combination) appears in a document. One general approach is to apply dimensionality reduction algorithms to represent the high-dimensional term vectors in a low-dimensional space. The dimensionality reduction can involve nonlinear projection methods such as Self-Organizing Maps (SOMs; Kohonen et al., 2000), linear projection methods such as Latent Semantic Analysis (Landauer & Dumais, 1997) or clustering models that characterize each document by a single latent cluster or topic (e.g. Popescul, Ungar, Flake, Lawrence, & Giles, 2000). As a result of the dimensionality reduction, neighboring points in the semantic space often represent words or documents with similar contextual usages or meaning.

* Tel.: +1 949 824 7642; fax: +1 949 824 2307.
E-mail address: mark.steyvers@uci.edu

0001-6918/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.actpsy.2009.10.010

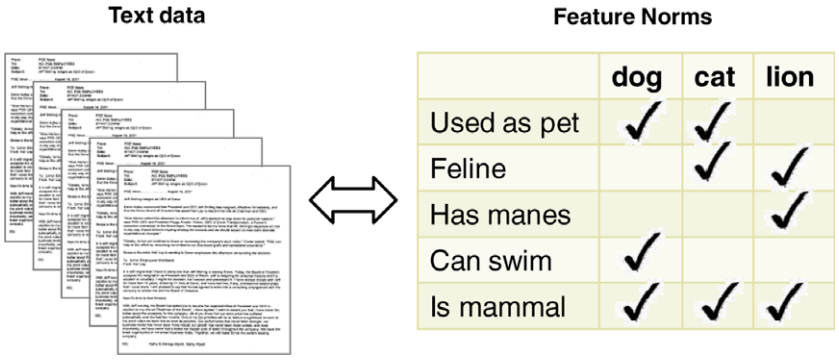

Fig. 1. Illustration of the overall goal of this research: combining statistical information text data and feature norming data.

These representations have been shown to model human knowledge in a variety of cognitive tasks (Landauer & Dumais, 1997).

A flexible unsupervised learning framework was recently introduced known as statistical topic models (Blei, Ng, & Jordan, 2003; Buntine & Jakulin, 2004; Griffiths & Steyvers, 2004; Griffiths, Steyvers, & Tenenbaum, 2007; Hofmann, 1999; Steyvers & Griffiths, 2007). The basic concept underlying topic modeling is that each document is composed of a probability distribution over topics, where each topic represents a probability distribution over words. The document-topic and topic-word distributions are learned automatically from the data and provide information about the semantic themes covered in each document and the words associated with each semantic theme. The underlying statistical framework of topic modeling enables a variety of interesting extensions to be developed in a systematic manner, such as correlated topics (Blei & Lafferty, 2006), hierarchical topic models (Blei, Griffiths, Jordan, & Tenenbaum, 2004; Li, Blei, & McCallum, 2007; Teh, Jordan, Beal, & Blei, 2006), time-dependent topics (Wang, Blei, & Heckerman, 2008), models that combine topics and syntax (Boyd-Graber & Blei., 2008; Griffiths, Steyvers, Blei, & Tenenbaum, 2005) as well as image features and text (Blei et al., 2003). Topic models have also been useful as cognitive models to explain human associations, gist extraction, and memory errors (Griffiths et al., 2007).

One drawback to this data-driven approach to semantic representation is that the resulting topics are not always easy to interpret. In addition, the topic representations become reliable only with large amounts of text data. Recently, topic models have been extended to incorporate background information in the form of human concepts from a thesaurus and ontologies from the world-wide web (Chemudugunta, Holloway, Smyth, & Steyvers, 2008; Chemudugunta, Smyth, & Steyvers, 2008a, 2008b; Steyvers, Chemudugunta, & Smyth, submitted for publication). This background knowledge can greatly improve the model when little text is available and facilitates the interpretation of learned semantic representations.

In this research, we propose to extend topic models with background knowledge in the form of feature norms (see Fig. 1). In these *feature-topic* models, the idea is that the presence of words in documents can be explained by both learned topics and predefined human knowledge about features. There are already some models that combine word co-occurrence information and featural information (e.g. Andrews, Vigliocco, & Vinson, 2005). One difference is that we will work with statistical topic models as the foundation for incorporating featural information. Also, in our model, we will treat features as latent causal factors that explain the presence of (some) words in documents. In contrast, the model by Andrews et al. (2005) treats both features and words as observed statistical information that is explained by latent clusters – therefore, features are not considered the underlying causal factors to explain word choices in documents. We will revisit the difference between these modeling approaches in a later section. In the present article, we will rely on the feature norms from De Deyne et al. (2008) and Ruts et al. (2004), henceforth called the Leuven Natural Concept Database.

(a)

document	missing word
The [ ] is related to the **pig** and they are both very fat. They both roll in the mud, and love water. The **pig** is also related to the [ ] because of the short tail. The difference is that the [ ] lives almost only in the wild and the **pig** lives on a **pig** farm. The [ ] looks a bit like the **rhinoceros** and the **elephant**, but they are not related. Because a **rhinoceros** has a horn and an **elephant** a trunk. And a [ ] lives mostly in water, and an **elephant** and **rhino** live on the savanna	*hippo*

(b)

pig, rhinoceros, elephant	*hippo*
boat, bus, tram, train	*airplane*
organ, piano, saxophone, trombone	*cello*
beaver, mouse, elephant, pig, toad, boat	*frog*
scissors, stick, tongs	*knife*

Fig. 2. (a) Example document where a single exemplar from the Leuven Natural Concept Database is missing and needs to be predicted. The missing word is *hippo*. Words in bold indicate observed exemplars from the Leuven Natural Concept Database, for which we have featural information available. (b) Example documents where words not part of the Leuven Natural Concept Database were removed and all word frequencies were set to one. Each italicized word shows the missing word that need to be predicted. The first document corresponds to the document shown in panel (a).

Fig. 2 motivates the development of the feature-topic models with the task of predicting the identity of missing words in documents. In Fig. 2a, a document[1] is shown where a single word is missing (the boxes hide repetitions of the same word). The missing word is part of the Leuven Natural Concept Database. The words in bold show the observed exemplars from the Leuven Natural Concept Database (*pig*, *elephant*, and *rhinoceros*). The words not in bold form the additional linguistic context. In our probabilistic framework, the goal is to develop models that give high posterior predictive probability to the missing word on the basis of the (probabilistic) representation given to the document. The missing word in Fig. 2a is *hippo* which can be predicted from a variety of sources of information, including features and the linguistic context. For example,

[1] The document is loosely based on a translation from a Dutch educational document from http://www.scholieren.com/werkstukken/21705. The document only is used for illustration purposes and was not part of the Dutch corpus.