Acta Psychologica xxx (2009) xxx–xxx

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 xxxx

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, this issue).

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

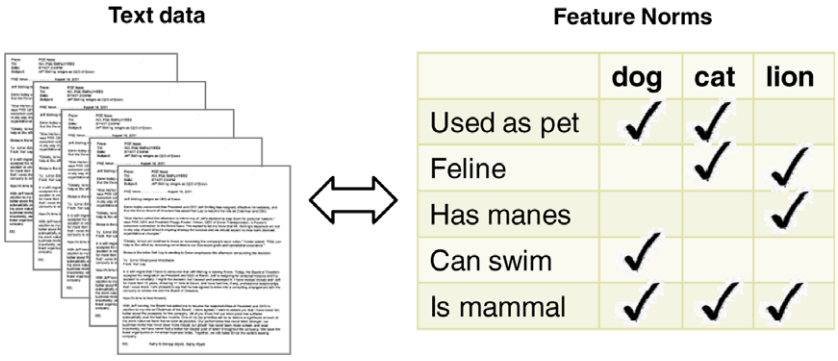

Fig. 1. Illustration of the overall goal of this research: combining statistical information text data and feature norming data.

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.

Please cite this article in press as: Steyvers, M. Combining feature norms and text data with topic models. *Acta Psychologica* (2009), doi:10.1016/j.actpsy.2009.10.010

one might infer the presence of a latent feature such as *thick skin*, *lives near water*, and *is mammal*. Such latent features could be used to explain why the words *pig*, *elephant*, and *rhinoceros* are present. In addition, it might be useful to utilize the linguistic context outside of the words in the feature norms. For example, the words *water*, *mud*, *savanna*, *wild*, and *fat* give additional cues about the identity of the missing word. This information could be part of a topic model if there is correlation across documents between the word *hippo*, and words such as *water*, *mud*, *savanna*, *wild*, and *fat*.

Fig. 2b illustrates other example documents (taken from a Dutch corpus explained in the next section) where all words that were not part of the Leuven Natural Concept Database were removed and all word counts were set to one. Therefore, the first line indicates a document where among other words, the exemplars *boat*, *bus*, *tram*, and *train* were present. The missing words that need to be predicted are shown in italics. These examples illustrate that for some of the documents, correct generalization requires the inference of a single underlying category (e.g. ***vehicles***) to predict the missing word (e.g. *airplane*). However, other generalizations might require more fine-grained featural information. For example, to predict *knife* from the observed words *scissors*, *stick*, and *tong*, it might be useful to infer the presence of a latent feature of *pointy things*.

The plan for the rest of the paper is as follows. Section 2 describes the feature norms and the Dutch corpus that will be used for all simulations. Section 3 reviews the basic principles of topic models and shows the results of applying a topic model to the Dutch corpus. Section 4 introduces the feature-topic model which combines features and topics into a single probabilistic model. Section 5 describes a series of experiments that evaluate the predictive performance of the topic and feature-topic models on the missing word task. In Sections 6 and 7, we conclude the paper with a brief discussion of related research, future directions and final comments.

2. Datasets

Our modeling work is based on two Dutch datasets: the Leuven Natural Concept Database (De Deyne et al., 2008; Ruts et al., 2004) and a large corpus of Dutch documents obtained from De Deyne (*personal communication*). Because both datasets are based on the Dutch language, this allows for direct comparisons between the statistical information contained in both text and feature norms (see Vandekerckhove, Verheyen, & Tuerlinckx, this issue, for another approach to extract statistical information from natural language corpora to predict aspects of semantic cognition).

2.1. Leuven Natural Concept Database

We used a version of the feature norms that included 129 exemplars from semantic categories including animals such as ***mammals***, ***birds***, ***reptiles***, and 166 exemplars from artifact categories such as ***vehicles***, ***musical instruments*** and ***clothing***. In total, we focused on 11 categories comprising 295 exemplars. In the feature norming study described by De Deyne et al. (2008), one group of subjects was asked to produce a list of features associated with each exemplar. Features could include perceptual or functional characteristics or any other background information that came to mind. For example, for the word *lion*, features produced included *has manes*, *is dangerous*, etc.[2] In the original norms, a separate matrix of the exemplar by features was maintained for the animal and artifact categories. In this research, we merged the two matrices leading to a single matrix. In the merging process, 28 features overlapped between the animal and artifact categories and there were 736 and 1267 unique animal and artifact features. This led to a matrix of 2031 features by 287 exemplars. Finally, we removed all features that related to only a single word leading to a final matrix of 1793 features by 287 exemplars. Each cell in this matrix contains the number of subjects who judged a particular feature to be applicable to a particular exemplar.

2.2. Dutch corpus data

The corpus was obtained from De Deyne (*personal communication*). The corpus contains web documents retrieved with a Dutch search engine by searching for a broad set of exemplars and retrieving a large number of documents for each exemplar of interest. This insures that infrequent exemplars in a category (e.g. *rhinoceros*) are represented in the corpus as well as the more frequent members of a category (e.g. *elephant*). The exemplars were taken from the Leueven concept database but also included a larger set of words from the Battig and Montague (1969) stimuli collected in Dutch by Storms (2001). For each exemplar, a large number of Dutch web documents were retrieved leading to a corpus of 540 M word tokens and 773 K documents (excluding non-Dutch language documents). We filtered this original set of documents by excluding all documents that did not contain any exemplar from the Leuven Natural Concept Database (note that the original corpus was constructed with a broader set of exemplars in mind). We also required that each document contained at least five occurrences of an exemplar, and between 50 and 500 words occurrences total. This excluded very short and very long documents. Finally, we excluded word types that occurred in fewer than 5 or more than 10 K documents. This led to a corpus of 21 M word tokens, 84 K documents, and a vocabulary of 69 K words. This corpus covers 287 of the 295 exemplars from the Leuven Natural Concept Database. Missing words included exemplars with special characters such as *pinguïn*, *t-shirt* and exemplars involving word combinations such as *engelse sleutel*, that present challenges for standard word-tokenizers. Finally, for the purpose of creating a test set to compare generalization performance of various topic models, we randomly sampled 5 K documents which had between 3 and 10 unique exemplars. A single exemplar was removed from each of these documents (and all word tokens associated with the exemplar were removed from the document) and was reserved for testing purposes. This led to a test set of 5 K missing words to predict. All remaining 21 M word tokens (regardless of whether the word tokens were part of documents that had or had not words removed) were used to train topic models.

3. Topic models

In this section, we begin with a brief review of probabilistic topic models and show some results of applying the topic model to a collection of Dutch documents. The topic model is a statistical learning technique for extracting a set of topics that describe a collection of documents. A topic t is represented by a multinomial distribution over the V unique word types in the corpus, $\boldsymbol{\varphi}^{(t)} = [\varphi_1^{(t)}, \ldots, \varphi_V^{(t)}]$ where $\varphi_w^{(t)} = p(w|t)$, and $1 \leqslant w \leqslant V$. Therefore, a topic can be viewed as a V-sided die and generating n word tokens from a topic is akin to throwing the topic-specific die n times. There are a total of T topics and the dth document is represented as a multinomial distribution over those T topics, $\boldsymbol{\theta}^{(d)} = [\theta_1^{(d)}, \ldots, \theta_T^{(d)}]$, where $\theta_t^{(d)} = p(t|d)$, and $1 \leqslant t \leqslant T$. The variables $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$ indicate which words are important for which topic and which topics are important for a particular document, respectively.

[2] In a separate experiment, a small number of subjects verified for each feature whether a particular exemplar contains that feature. This insured that some features that were not generated for a particular exemplar might nevertheless be judged relevant. For example, is mammal might not be produced as a feature for the exemplar lion but it will be checked as a feature in the verification task. The feature verification stage improves the quality of the feature norms compared to feature norms that do not rely on this second stage process.

Please cite this article in press as: Steyvers, M. Combining feature norms and text data with topic models. *Acta Psychologica* (2009), doi:10.1016/j.actpsy.2009.10.010

Generating a word token for a document d involves first selecting a topic t from the document-topic distribution $\theta^{(d)}$ and then selecting a word from the corresponding topic distribution $\varphi^{(t)}$. This process is repeated for each word token in the document. Let $\boldsymbol{z}$ be the random variable that represents the topic indices sampled from $\theta^{(d)}$. We write $p(z_i = t|d)$ as the probability that the tth topic was sampled for the ith word token (in document d) and $p(w_i|z_i = t)$ as the probability of word w_i under topic t. The model specifies the following conditional probability of the ith word token in a document:

$$p(w_i|d) = \sum_{t=1}^{T} p(w_i|z_i = t)p(z_i = t|d) \quad (1)$$

Fig. 3, top panel, illustrates the generative process of the topic model with a toy example involving two example topics and five words in the vocabulary. The left panel shows two example topics, that give high probability to the words *money*, *loan*, *bank*, and *river*, *stream*, *bank*, respectively. Note how the same word *bank* can occur in multiple topics – there is no restriction in the topic model that words are only represented in a single topic. The left panel also shows how three example documents can be generated by either selecting word tokens from a single topic or by mixing the two topics in different proportions. The superscript numbers next to the word tokens in the documents represent the particular topic that was chosen to generate the word token.

In the latent Dirichlet allocation model (Blei et al., 2003; Griffiths & Steyvers, 2004), Dirichlet priors are placed on both θ and φ in order smooth the word-topic and topic-document distributions. In many applications, a symmetric Dirichlet density with single hyperparameters α and β are used for θ and φ, respectively. However, in this research, we will use an asymmetric Dirichlet prior on θ where each topic has a prior weight α_j that determines how likely topic j is to be sampled across the whole corpus.

The sequential process of first picking a topic from a topic distribution, and then picking a word from a word distribution associated with that topic can be formalized as follows:

1. For each topic $t \in \{1, \ldots, T\}$, select a word distribution $\boldsymbol{\phi}^{(t)} \sim \text{Dirichlet}(\beta)$
2. For each document $d \in \{1, \ldots, D\}$
   (a) Select a distribution over topics $\theta^{(d)} \sim \text{Dirichlet}(\boldsymbol{\alpha})$
   (b) For each word token i in document d
      i. Select a topic $z_i \sim \text{Discrete}(\theta^{(d)})$
      ii. Generate a word token from topic z_i, $w_i \sim \text{Discrete}(\boldsymbol{\phi}^{(z_i)})$

This generative process can be summarized by the graphical model shown in Fig. 4a (see Griffiths, Kemp, and Tenenbaum (2008), and Lee (2008) for other graphical model examples in the area of cognitive science). In this graphical notation, shaded and unshaded variables indicate observed and latent (i.e., unobserved) variables, respectively, and the arrows indicate the conditional dependencies between variables. The plates (the boxes in the figure) refer to repetitions of sampling steps with the variable in the right corner referring to the number of samples. For example, the inner plate over z and w illustrates the repeated sampling of topics and words until N_d words have been generated for document d. The plate surrounding θ illustrates the sampling of a distribution over topics for each document d for a total of D documents. The plate surrounding φ illustrates the repeated sampling of word distributions for each topic until T topics have been generated.

Given the words in a corpus, the inference problem involves estimating the word-topic distributions φ, the topic-document distributions θ, and the topic assignments $\boldsymbol{z}$ to word tokens. Fig. 3, bottom panel, illustrates the problem of statistical inference. The only observed variables are the word occurrences in documents and all latent variables are indicated by question marks.

The latent variables can be estimated in a completely unsupervised manner without any prior knowledge about topics or which topics are covered by what documents. One efficient technique for obtaining estimates is through Markov Chain Monte Carlo (MCMC) using collapsed Gibbs sampling (Griffiths & Steyvers, 2004). In this approach, estimation is done only on the z assignments of word tokens to topics, while integrating out ("collapsing") the remaining latent variables. Words are initially assigned randomly to topics and the algorithm then iterates through each word in the corpus and samples a topic assignment given the topic assignments of all other words in the corpus. This process is repeated until a steady state is reached and the topic assignments to words are then used to estimate the word-topic and topic-document distributions. For more information on the Gibbs sampling

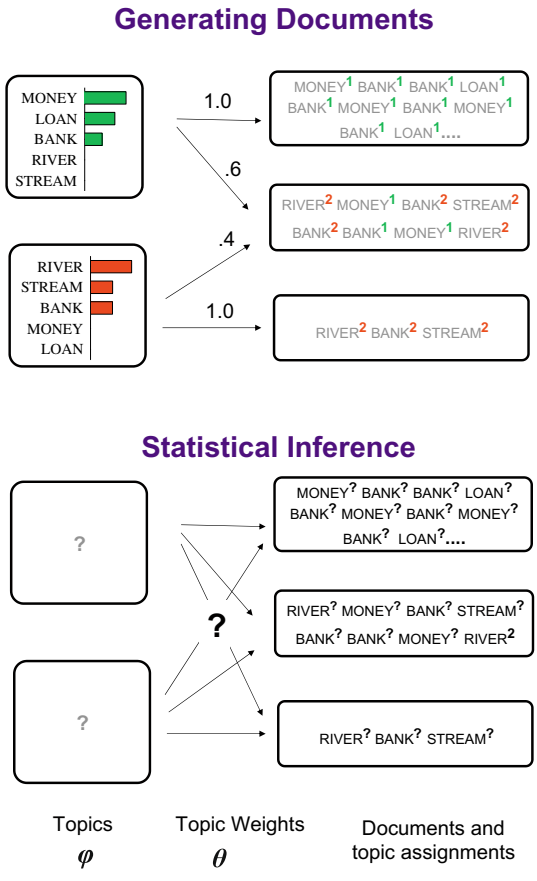

Fig. 3. Illustration of the generative process to sample words in document (top panel) and the problem of statistical inference (bottom panel).

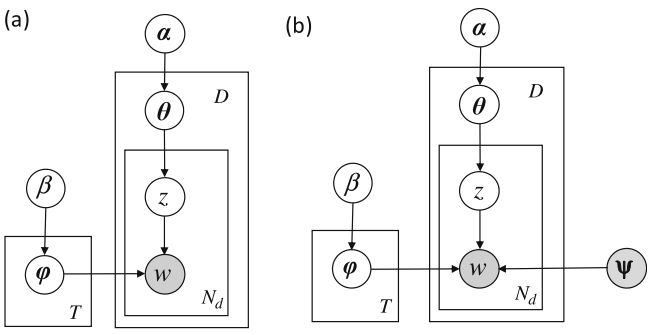

Fig. 4. Graphical models for the topic model (a), and the feature-topic model (b).

Please cite this article in press as: Steyvers, M. Combining feature norms and text data with topic models. *Acta Psychologica* (2009), doi:10.1016/j.actpsy.2009.10.010

process, please see an introductory article by Steyvers and Griffiths (2007).

To summarize, the topic model gives several pieces of information that are useful for understanding documents. The topic-document distributions indicate the important topics for a particular document. The word-topic distributions indicate which words are important for which topic. Finally, the probabilistic assignments of word tokens to topics are useful for tagging and word-sense disambiguation; this gives information about the role the word is playing in a specific document context and can help to disambiguate multiple meanings of a word.

3.1. Applying the topic model to Dutch corpus data

We ran the topic model on the Dutch corpus data with $T = 10$, 20, 50, 100, 200, and 400 topics. The hyperparameter β was set to 0.1. One difference between the standard topic model and the topic model pursued in this research is that we used an asymmetric Dirichlet prior on θ, with a vector $\boldsymbol{\alpha}$ containing hyperparameter values for every topic. This is useful when some topics are expressed in many or just a few documents across the collection. With an asymmetric prior, more skewed marginal distributions over θ can be obtained and each hyperparameter α_i indicates which topics are important across the whole corpus. Instead of estimating each individual α_i value through MCMC sampling methods, we optimized them using fixed point update equations (Minka, 2000; Wallach, 2006). See also Appendix A of Chemudugunta et al. (2008b) for more details.

For each number of topics, we ran five different MCMC chains (i.e., starting the Gibbs sampler five times with different random seeds) and ran each chain for 1000 iterations. At the end of this process, we took a single sample from each chain. While it is often difficult to assess convergence of MCMC samplers, we know from previous work that several hundred iterations are usually sufficient to lead to asymptotic performance on a variety of generalization tasks (e.g. Rosen-Zvi, Chemudugunta, Griffiths, Smyth, & Steyvers, 2010).

Fig. 5 shows three example topics (arbitrarily renumbered to topics 1, 2, and 3) from a single Gibbs sample with $T = 100$ topics. Each topic is illustrated by the 19 words that have the highest probability under that topic. English translations of the Dutch words are provided to facilitate the interpretation of the topic. The topics highlight semantic themes related to boating, horses and African animals.

To better understand the statistical representation of exemplars from the Leuven Natural Concept Database, we analyzed all pairwise dissimilarities between exemplars. The dissimilarity between two words can be calculated based on the distributional differences between $p(z|w_1)$ and $p(z|w_2)$, the conditional topic distributions for two words w_1 and w_2. This difference measures the extent to which two words share the same topics. We used the symmetrized Kullback Leibler (KL) divergence $S(p, q)$ to measure the distributional dissimilarity between two distributions p and q, where $S(p,q) = 1/2[D(p,q) + D(q,p)]$ and $D(p,q) = \sum_{j=1..T} p_j \log_2 p_j/q_j$, the KL divergence between two distributions. The dissimilarity measure was averaged over the five samples that were extracted for each simulation.

Fig. 6 shows the pairwise dissimilarities of all 287 exemplars from the Leuven Natural Concept Database with 10 topics (panel a) and 200 topics (panel b). For visual clarity, the individual words are not listed. Instead, the words are grouped by the corresponding categories (predefined in the Leuven Natural Concept Database) and the dashed lines separate the different categories. The gray values in the matrix represent the symmetrized KL divergences with darker shades representing higher degrees of topic similarity. Fig. 6 shows that with only 10 topics, the model discriminates mostly between living concepts and artifacts. However, a few individual categories such as musical instruments and clothing are separated from other categories. With 200 topics, the model separates most categories from each other. In addition, for some categories, such as vehicles, weapons and tools, the model does not appear to represent the category at all, making words within these categories as dissimilar to each other as to words outside these categories. Overall, these results indicate that the model dimensionality influences the degree to which words, categories, and superordinate categories can be distinguished from each other.

4. Feature-topic model

The feature-topic model is an extension to the topic model where we add the featural information from the semantic feature norms to the model. Each feature is treated as an additional "topic" in the model that cannot be modified by the learning algorithm. The idea is that the generative process of creating a document can be based on a variety of sources of information, including featural information. Therefore, when we add F features to the T topics of the topic model, this results in an effective set of $T + F$ topics for each document. For example, when the theme of a document is related to African animals, one can imagine that a feature such as *lives in Africa* can explain some of the words in a document. This is especially useful when there is not enough statistical information in the word occurrences across documents to infer a similar topic that would put all probability mass on words related to *lives in Africa*.

word	$\varphi^{(1)}$	word
boot	0.076	boat
zee	0.031	sea
schip	0.019	ship
meter	0.018	meter
zwemmen	0.017	swimming
dolfijn	0.016	dolphin
haai	0.016	shark
eiland	0.015	island
varen	0.014	sail
strand	0.013	beach
boten	0.011	boats
boord	0.011	board
dolfijnen	0.010	dolphins
walvis	0.009	whale
haven	0.009	harbor
vaart	0.008	channel
duiken	0.008	dive
kust	0.008	coast
schepen	0.007	ships

word	$\varphi^{(2)}$	word
paard	0.141	horse
paarden	0.049	horses
haan	0.024	rooster
ezel	0.024	mule
pony	0.016	pony
zadel	0.014	saddle
rijden	0.013	ride
kar	0.011	cart
stal	0.011	stable
ruiter	0.008	horseman
deken	0.007	blanket
paardrijden	0.006	horse riding
manege	0.006	riding-school
zweep	0.005	whip
ezels	0.005	donkeys
wei	0.005	meadow
koets	0.005	carriage
veulen	0.005	foal
benen	0.004	legs

word	$\varphi^{(3)}$	word
olifant	0.091	elephant
Afrika	0.053	Africa
zebra	0.038	zebra
olifanten	0.032	elephants
krokodil	0.030	crocodile
Afrikaanse	0.028	African
zuid	0.021	south
giraf	0.020	giraffe
nijlpaard	0.019	hippo
giraffe	0.012	giraffe
dieren	0.012	animals
slurf	0.010	trunk
krokodillen	0.009	crocodiles
luipaard	0.008	leopard
safari	0.007	safari
leeuwen	0.007	lions
Kenia	0.007	Kenya
park	0.006	park
neushoorn	0.006	rhinoceros

Fig. 5. Three examples topics, numbered as topics 1, 2, and 3. For each topic, the words are sorted by $\varphi^{(j)} = p(w|z = j)$, the probability of a word in topic j, and the top 19 Dutch words are shown in the left columns. English translations are provided in the right columns.

Please cite this article in press as: Steyvers, M. Combining feature norms and text data with topic models. *Acta Psychologica* (2009), doi:10.1016/j.actpsy.2009.10.010

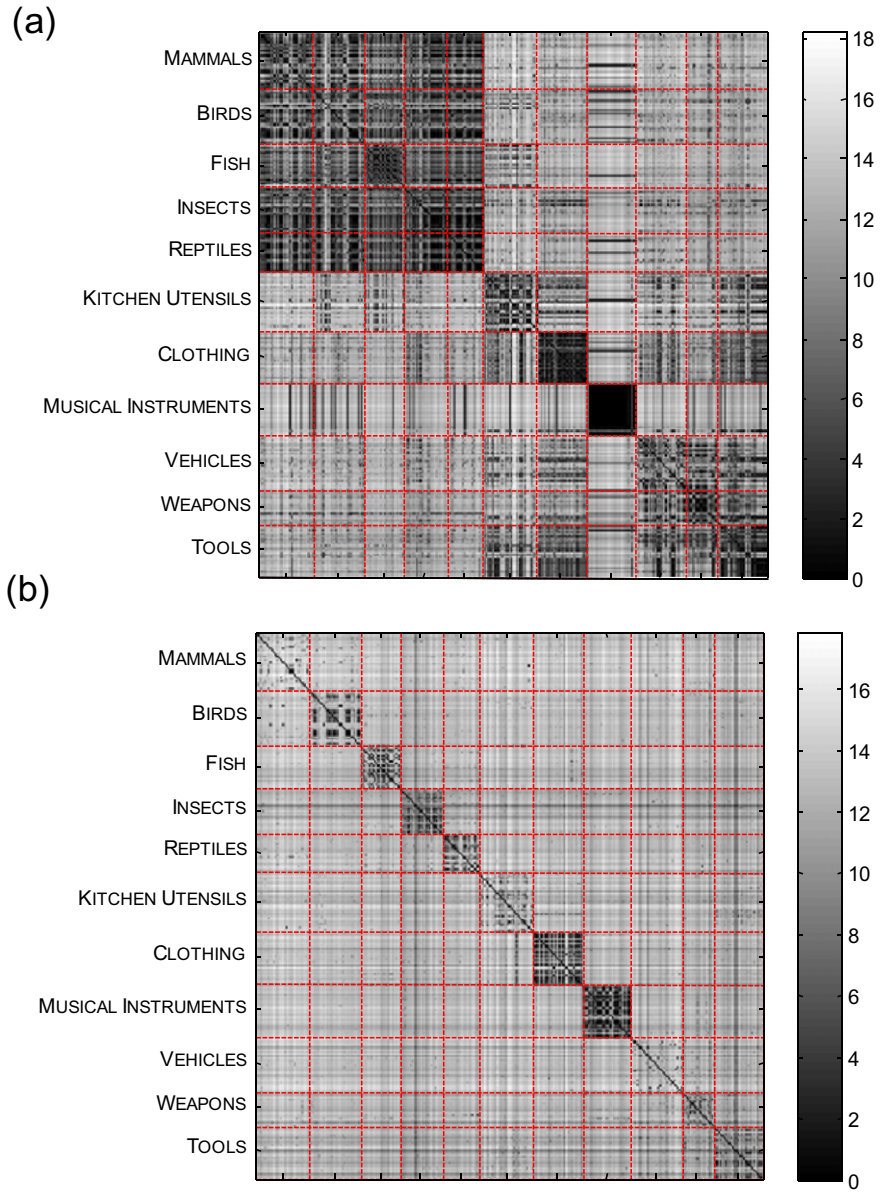

Fig. 6. The pairwise (dis)similarities between 287 keywords from the Leuven Natural Concept Database as measured by the symmetrized KL distance between topic distributions for 10 topics (panel a) and 200 topics (panel b). The legend on the right-hand side shows how the gray values map to the KL distances – darker values indicate smaller distances and therefore greater overlap in topic distributions for two words.

In this model, we assume that each feature is associated with a distribution of words. Fig. 7a illustrates the data we have available from the Leuven Natural Concept Database. Originally, the feature by exemplar matrix contains the number of subjects who judge a feature to be associated with a particular exemplar. In order to convert this to word distributions, we normalize each row in the matrix and treat each feature f as a multinomial distribution $\boldsymbol{\psi}^{(f)} = [\psi_1^{(f)}, \ldots, \psi_V^{(f)}]$ where $\psi_w^{(f)} = p(w|f)$, and $1 \leqslant w \leqslant V$. This information can be directly utilized as a prelearned topic in the topic model. For example, Fig. 7b shows the word distribution for the feature "has manes". In this feature, non-zero probability goes to *lion*, *zebra*, *donkey*, and *horse*. The fact that words can have different (non-zero) probabilities reflects the uncertainty that subjects have about the presence of some features, such as the presence of manes for donkeys.

The feature-topic model is simply an extension of the LDA model where we have a number of learned topics as well as constrained topics based on features. This model is similar to the concept-topic model that was recently applied to concepts from a thesaurus (Steyvers, Chemudugunta, Smyth, submitted for publication). In the concept-topic model, additional background information is provided by concepts defined by lexicographers in a thesaurus. In the feature-topic model, the background information is provided by human feature norms instead.

In the feature-topic model, the conditional probability of the ith word w_i given a document d is,

$$p(w_i|d) = \sum_{j=1}^{T+F} p(w_i|z_i = j)p(z_i = t|d) \qquad (2)$$

where the indices $1 \leqslant j \leqslant T$ refer to all learned topics and indices $T+1 \leqslant j \leqslant T+F$ refer to all features. In this generative process, a topic j is sampled from the distribution over topics and features for the particular document. If $j \leqslant T$, a word is sampled from the word-topic distribution $\boldsymbol{\varphi}^{(j)}$ and if $T+1 \leqslant j \leqslant T+F$, a word is sampled from the word-feature distribution $\boldsymbol{\psi}^{(f)}$ where $f = j - T$.

Please cite this article in press as: Steyvers, M. Combining feature norms and text data with topic models. *Acta Psychologica* (2009), doi:10.1016/j.actpsy.2009.10.010

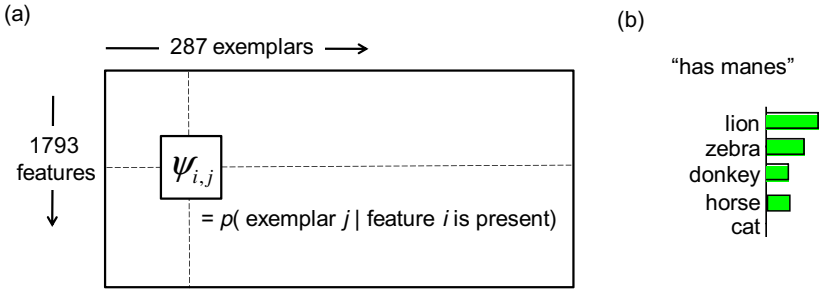

Fig. 7. Illustration of the feature by exemplar matrix extracted from the feature verification task from the Leuven Natural Concept Database. The rows in this matrix are normalized such that each entry is the conditional probability of a word given that a feature is present (a). An example word distribution associated with the feature *has manes* (b).

The topic model can be viewed as a special case of the feature-topic model when there are no features present, i.e. when $F = 0$. At the other extreme of this model where $T = 0$, the model relies entirely on predefined word distributions associated with the features.

The complete generative process can be described as follows:

1. For each topic $t \in \{1, ..., T\}$, select a word distribution $\phi^{(t)} \sim \text{Discrete}(\beta_\phi)$
2. For each feature $f \in \{1, ..., F\}$, associate a predefined word distribution $\psi^{(f)}$
3. For each document $d \in \{1, ..., D\}$
   (a) Select a distribution over topics and features $\theta^{(d)} \sim \text{Discrete}(\alpha)$
   (b) For each word token i in document d
      i. Select a component $z_i \sim \text{Discrete}(\theta^{(d)})$
      ii. If $z_i \leqslant T$, generate a word from topic z_i, $w_i \sim \text{Discrete}(\phi^{(z_i)})$; otherwise, generate a word from feature $f = z_i - T$, $w_i \sim \text{Discrete}(\psi^{(f)})$

Fig. 4b shows the corresponding graphical model. As in the previous topic modeling, we used an asymmetric Dirichlet prior on θ, with a vector α containing hyperparameter values for every topic and feature. This flexibility in the model is especially important when incorporating features into the model – features are not the same as learned topics. There is no opportunity to change the word distributions associated with features so a priori, it is not clear how useful each feature will be. For example, many of the features are highly idiosyncratic (e.g. "occasionally occur in films" and "makes an irritating sound") and are unlikely to be useful in explaining word occurrences. In the model, such features can be given a low weight in the Dirichlet prior, making it unlikely that such a feature is sampled for any of the documents.

4.1. Applying the feature-topic model to Dutch corpus data

To apply this model to corpus data, an important issue is to decide on the vocabulary of words. In the previous section, we applied the topic model to a 69 K word vocabulary. In the feature norms, the features for only 287 words were verified and the applicability of the features for other words is unknown. Therefore, we have no direct observations about the probabilities of a large number of words in each word-feature distribution. One possibility is to treat these probabilities as latent variables to be estimated by the model. However, such a model would be non-trivial to estimate and is beyond the scope of this paper. Instead, we chose to restrict the vocabulary in our simulations to just the 287 words in the feature norms and removed all other words from the corpus. Although this is a rather dramatic change in vocabulary size, note that in the generalization task to be described below, all models are tested on the *same* task involving the prediction of missing words from the 287 words that are known to the topic models as well as the feature-topic models. In addition, we can compare the performance of the feature-topic model with a topic model that also is restricted to the 287 word vocabulary.

As mentioned before, we inferred all the latent variables in the model through collapsed Gibbs sampling and optimized the hyperparameters α using fixed point update equations. We ran the model on the Dutch corpus data with $T = 10, 20, 50, 100$, and 200 topics. The hyperparameter β was set to 0.1. For each number of topics, we ran five different MCMC chains and extracted a single sample after running each chain for 1000 iterations.

Fig. 8 illustrates the variability in the estimated hyperparameters for each feature. The results are taken from a topic model with no learned topics ($T = 0$) and the hyperparameters are averaged over samples. A large fraction of features are associated with a low prior weight and are essentially ignored as potential mixture components in documents. Only a small number of features are used to explain word occurrences in documents. To better understand which features receive a low or high prior weight, we plotted the hyperparameters against the size of the feature, which we define as the number of words that have a non-zero probability for the feature. For example, the feature *has manes* has size 4 because four exemplars have non-zero probability. Fig. 8 shows that features associated with few words (e.g. *used to bake*) are estimated to be a priori more useful than features that are associated with many words (e.g. *has round eyes*).

This pattern is related to the *size principle* in feature models of similarity that emerges from a Bayesian analysis of induction prob-

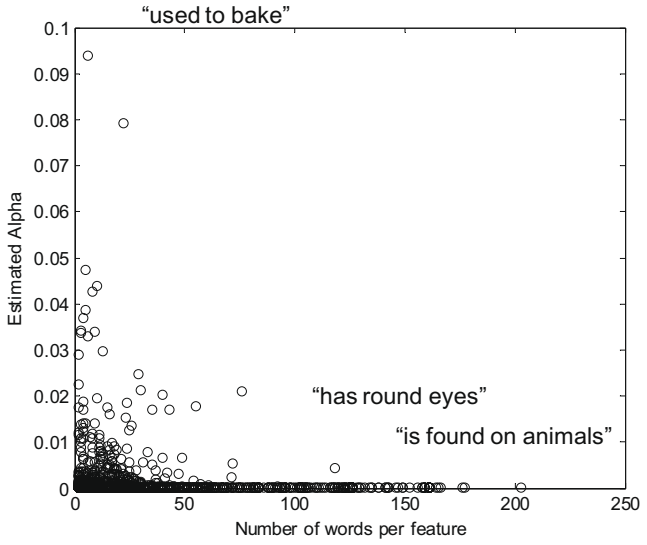

Fig. 8. The estimated α_i parameters for a feature-topic model with no learned topics ($T = 0$) as a function of the number of words with non-zero probability in that feature.

Please cite this article in press as: Steyvers, M. Combining feature norms and text data with topic models. *Acta Psychologica* (2009), doi:10.1016/j.actpsy.2009.10.010

lems (Navarro & Perfors, this issue; Tenenbaum & Griffiths, 2001). According to this principle, a feature should be weighted inversely proportional to the number of exemplars that are associated with the feature. This can lead to feature models that can account for human similarity ratings (e.g. Tenenbaum & Griffiths, 2001) and more generally to optimal representations (Navarro & Perfors, this issue). In our case, we observe that the model is approximately implementing a size principle through the learned hyperparameter weights on the features. By giving larger weights to features associated with few words, the model can best explain the observed word occurrences. This suggests that the size principle for features is not only useful to account for similarity ratings for also for corpus data.

5. Comparing models on the missing word task

To assess the topic and feature-topic models quantitatively, we evaluate their capability to generalize to unobserved words in documents. Models that are able to represent the latent content of a document with features and/or learned topics should be able to extrapolate from the latent content to new words. In the Dutch corpus, we created a test set where in each of 5 K documents one of the words, an exemplar from the Leuven Natural Concept Database, was removed (and all word tokens corresponding to this word were removed). The test documents were chosen such that there were always between 2 and 9 observed exemplars in each document in addition to any others words not part of the Leuven Natural Concept Database.

In our simulations, we compared three models, a feature-topic model with a vocabulary of 287 words, a topic model with the same limited vocabulary of 287 words, and a topic model with an extended vocabulary of 69 K words. All feature-topic models were run on the limited vocabulary of the 287 words that are part of the Leuven Natural Concept Database. While this removes much potentially useful information about the content of documents, we can compare this model that utilizes featural information with a topic model that uses the same limited vocabulary but does not utilize the featural information. This comparison allows us to assess the benefit of the featural information. We can also compare the topic model with the limited vocabulary (and no features) with the topic model that uses the extended vocabulary. This comparison allows to assess the amount of information present in the extra linguistic context provided by the words that are not in the Leuven Natural Concept Database. An additional useful comparison model would be a feature-topic model that is run on the extended vocabulary. However, as mentioned earlier, this requires specifying a model for extrapolating featural information to words not part of the feature norms. This model is outside the scope of the current article. Therefore, in our evaluation, we can assess the benefit of featural information or additional linguistic context, but not in combination.

For the topic and feature-topic model, we used Eqs. (1) and (2), respectively, to calculate the posterior predictive distribution over words in test documents. This is the distribution over words when one uses the latent set of topics and/or features in a forward generative process to generate words. The predictive distribution was averaged over samples. Because the nature of the task is to predict a single word that is not currently present in a document, we restricted the predictive distribution to the words in 287 word vocabulary that are not part of the current document. This is an important control – the models in the comparisons differ in the size of the vocabulary but we compared all models on the same limited vocabulary. We investigated several probability and ranking-based measures for performance but will focus here on a single measure only, the median percentile rank. The percentile rank of the missing word is based on the rank in the predictive distribution expressed in a percentage of the total number of words. Then we take the median of all percentile ranks across the 5 K test documents.

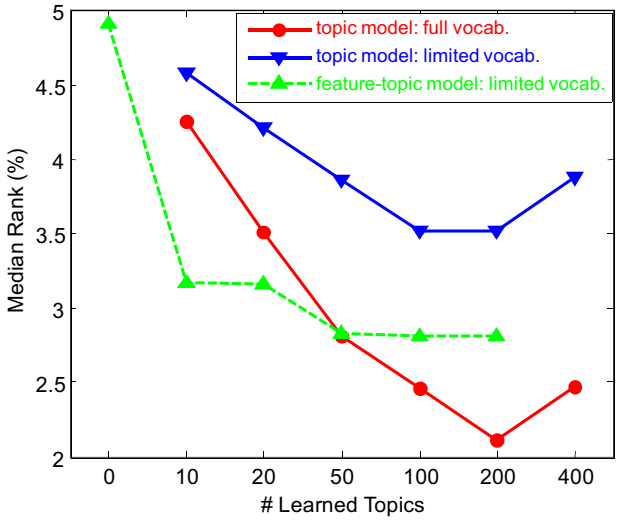

Fig. 9. Results of the generalization task. The median percentile rank is plotted for three topic models as a function of the number of learned topics.

Fig. 9 shows the median percentile rank for the three models as a function of the number of learned topics. Note that when the number of learned topics equals zero, the feature-topic model still relies on all the fixed topics derived from the feature norms. The baseline model for comparison purposes is the topic model with the restricted vocabulary. The figure shows that for all model dimensionalities, adding featural information improves performance. This result is important because in the model, a feature is represented in the same way as a topic. Both are treated as probability distributions over words. In principle, any of the fixed topics determined by the feature norms can be mimicked by a learned topic. However, because the performance of a feature-topic model outperforms the topic model (on the same vocabulary), it shows that although the models can mimic each other, a topic model is not able to extract the same featural information from a text corpus as is present in the feature norms.

Fig. 9 also shows that the performance is clearly affected by the number of learned topics. For the feature-topic model, performance is worst when there are no learned topics. Therefore, having learned topics that adapt to the corpus helps performance above and beyond the featural information. Adding too many learned topics however might negatively impact performance. For the topic model, adding 400 learned topics leads to worse generalization performance indicating that the model is starting to overfit the data.

Finally, Fig. 9 shows that larger vocabularies improve generalization performance, even when the generalization test is performed on the same limited vocabulary. A topic model with a 69 K word vocabulary learns correlations between words that are helpful to generalize to the limited set of 287 words. Therefore, the additional linguistic context improves performance. Overall, the results show that text models can be enhanced by adding prior featural information, learned topics as well as additional linguistic context.

6. Related work

The approach by Andrews et al. (2005) is most similar to our modeling work. They propose a model that combines features with

Please cite this article in press as: Steyvers, M. Combining feature norms and text data with topic models. *Acta Psychologica* (2009), doi:10.1016/j.actpsy.2009.10.010

word-document co-occurrence information in a probabilistic mixture model. In their model, the words in documents as well as the associated features of words are observed variables that are explained by a generative process. In other words, the features are treated as random variables that are affected by latent factors. In contrast, in our modeling approach, we turn the cause-effect relation around – the features are treated as latent random variables that are the causal factors that explain the presence of the observed words. Another key difference is that the mixture model by Andrews et al. assumes that each document is generated by a single latent cluster. This makes the model similar to many clustering models (e.g. Cutting, Karger, Pedersen, & Tukey, 1992; Popescul et al., 2000). The drawback of these methods is that documents that cover a diverse set of topics can only be represented by a single cluster leading to problems in interpretation (e.g. Newman, Chemudugunta, Smyth, & Steyvers, 2006). Also, it has been shown that topic models which allow multiple latent factors per document outperform clustering models in generalization tasks (Blei et al., 2003).

Zeigenfuse and Lee (2008, this issue) also proposed a method for finding the set of important features in the feature norms. In their probabilistic approach, they automatically select the set of features on the basis of human similarity data. Because similarity data are based on pairwise similarity ratings between words, it leads to a focus on features that best explains pairwise relations between words. In our work, we also used an external source of information to constrain features but instead of similarity data, we utilized text data. The features that receive high weight in the model are those that can explain word occurrences in documents. Because there are typically many words in a document, the model tries to explain the *setwise* relations between observed words.

7. Discussion and conclusion

We have proposed a probabilistic framework for combining data-driven topics and features provided by semantic feature norms. We introduced the feature-topic model, which is a straightforward extension of the topic model, to utilize semantic features in a topic modeling framework. This model represents documents as a mixture of learned topics and fixed word distributions derived from feature norms. Our experimental results showed that incorporating featural information into a statistical model for text improves generalization performance. We also showed that having learned topics that adapt to the linguistic context of a corpus helps performance above and beyond the featural information. One reason for this is that the model learns "ad hoc" categories of words (cf. Barselou, 1983) that explain relations between words that would be difficult to capture by featural information alone. For example, the leftmost topic in Fig. 5 groups words together based on a boating theme. This connects words such as *boats*, *dolphins*, and *sharks* that are unlikely (but not impossible) to be connected through features. Similarly the middle topic in Fig. 5 relates to a theme of horse riding and connects words such as *horse* and *whip*. It is not clear what feature would connect these words. Of course, one can imagine a relation such as *a whip can be used to prod a horse*, but the feature norms were not designed to extract such asymmetric relations between items. The features norms are limited to propositions involving a single argument such as *horse is a mammal*. Other knowledge databases such as ConceptNet (Havasi, Speer, & Alonso, 2007) however have been designed to extract multi-argument propositions from humans. In the absence of such knowledge bases, the learned topics in the feature-topic model can fill in the gaps of the feature norms and are able to connect words related to ad hoc categories.

The joint modeling of text and featural information reveals some other interesting differences between the statistical information contained in text and feature norms. First, not all features that connect words might be mentioned in feature production norms because features are produced with a single concept in mind, and not a set of concepts and their common features. Another difference is that the feature norms were designed with only a single meaning of each exemplar in mind, therefore facilitating the representation of concepts. However, many of the exemplars in the feature norms have multiple meanings. For example, the Dutch word *weegschaal* was intended as the English word *scale* in the category of kitchen utensils and appliances. However, in Dutch, it can also refer to a horoscope sign. Therefore, one can imagine that in one context the words *weegschaal* and *leeuw* (*lion*) are related because both are signs, but quite dissimilar in another context. Topic models are able to represent such ambiguities in meaning by the uncertainty in the topic distribution (Griffiths et al., 2007).

We view the current feature-topic models as a starting point for exploring more expressive generative models that can potentially have wide-ranging applications. One obvious extension of the model would be to allow the discovery of novel members of a feature. For example, if a feature such as *lives in Africa* has current members such as *lion*, *rhinoceros*, *zebra* and there are many co-occurrences of these words in documents with the novel word *aardvark*, it might be reasonable to infer that *aardvark* is also a member of the feature. Concept-topic models (Steyvers et al., submitted for publication) were similarly developed to discover new members of a concept. Another direction of research is to automatically extract assertions about concepts and features directly from text. There is already some work in this area mostly using hypothesis testing methods (e.g. Baroni & Lenci, 2008). We expect that an explicit generative framework built on feature-topic models will prove to be useful when utilizing the linguistic context and assertions about concepts and features stated in text to automatically populate feature norms with new concepts.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation under Award Number IIS-0083489. We are extremely grateful to one anonymous reviewer, and to Dan Navarro and Simon Dennis, for their very helpful comments on an earlier version of this paper.

References

Andrews, M., Vigliocco, G., & Vinson, D. (2005). The role of attributional and distributional information in semantic representation In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the twenty seventh annual conference of the cognitive science society*.

Baroni, M., & Lenci, A. (2008). Concepts and properties in word spaces. In Alessandro Lenci (Ed.), *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science* (special issue of the *Italian journal of Linguistics*, *20*(1), 55–88).

Barselou, Lawrence. W. (1983). Ad hoc categories. *Memory and Cognition, 11*, 211–227.

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the connecticut category norms. *Journal of Experimental Psychology Monographs, 80*, 1–45.

Blei, D., & Lafferty, J. (2006). Correlated topic models. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems* (Vol. 18, pp. 147–154). Cambridge, MA: MIT Press.

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems* (Vol. 16). Cambridge, MA: MIT Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Boyd-Graber, J., & Blei, D. (2008). Syntactic topic models. In *Neural information processing systems*. (Vol. 21, pp. 185–192). Cambridge, MA: MIT Press.

Buntine, W. L., & Jakulin, A. (2004). Applying discrete PCA in data analysis. In M. Chickering & J. Halpern (Eds.), *Proceedings of the 20th conference on uncertainty*

in artificial intelligence (pp. 59–66). San Francisco, CA: Morgan Kaufmann Publishers.

Chemudugunta, C., Holloway, A., Smyth, P., & Steyvers, M. (2008). Modeling documents by combining semantic concepts with unsupervised statistical learning. In *Proceedings of the 7th international semantic web conference* (pp. 229–244). Berlin: Springer Verlag.

Chemudugunta, C., Smyth, P., & Steyvers, M. (2008a). Combining concept hierarchies and statistical topic models. In *ACM 17th conference on information and knowledge management*.

Chemudugunta, C., Smyth, P., & Steyvers, M. (2008b). Text modeling using unsupervised topic models and concept hierarchies. *Technical report*, url: <http://arxiv.org/abs/0808.0973>.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Memory, 8*(2), 240–248.

Cutting, D. R., Karger, D., Pedersen, J. O., & Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the fifteenth annual international ACM SIGIR conference on research and development in information retrieval* (pp. 318–329). New York, NY: ACM Press.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., et al. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods, 40*(4), 1030–1048.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modeling* (pp. 59–100). Cambridge, MA: Cambridge University Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science, 101*, 5228–5235.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. *Advances in neural information processing* (Vol. 17). Cambridge, MA: MIT Press.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. T. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244.

Havasi, C. H., Speer, R., & Alonso, J. (2007). ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In *Proceedings of recent advances in natural language processing.*

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual ACM conference on research and development in information retrieval* (pp. 50–57). New York, NY: ACM Press.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., et al. (2000). Self organization of a massive document collection. *IEEE Transactions on Neural Networks (Special Issue on Neural Networks for Data Mining and Knowledge Discovery), 11–3*, 574–585.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin and Review, 15*(1), 1–15.

Li, W., Blei, D., & McCallum, A. (2007). Nonparametric Bayes Pachinko allocation. In *Conference on uncertainty in artificial intelligence* (UAI).

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*, 547–559.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology, 126*(2), 99–130.

Minka, T. P. (2000). Estimating a Dirichlet distribution. *Technical report*. Massachusetts Institute of Technology.

Navarro, D. J. & Perfors, A. F. (this issue). Similarity, feature discovery and the size principle. *Acta Psychologica*, doi:10.1016/j.actpsy.2009.10.008.

Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. In *Springer lecture notes in computer science (LNCS) series – IEEE international conference on intelligence and security informatics*

Popescul, A., Ungar, L. H., Flake, G. W., Lawrence, S., & Giles, C. L. (2000). Clustering and identifying temporal trends in document databases. In *Proceedings of the IEEE advances in digital libraries* (pp. 173–182). Los Alamitos, CA: IEEE Computer Society.

Rogers, T., & McClelland, J. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge: The MIT Press.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems, 28*(1).

Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Flemish norm data for 13 natural concepts and 343 exemplars. *Behavior Research Methods, Instruments, and Computers, 36*, 506–515.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review, 81*(3), 214–241.

Steyvers, M., Chemudugunta, C., & Smyth, P. (submitted for publication). Combining background knowledge and learned topics. *Topics in Cognitive Science* (topiCS).

Steyvers, M., & Griffiths, T. L. (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis*. Erlbaum.

Storms, G. (2001). Flemish category norms for exemplars of 39 exemplars: A replication of the Battig and Montague (1969) category norms. *Psychologica Belgica, 41*, 145–168.

Teh, Y. W., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association, 101*(476), 1566–1581.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*, 629–641.

Vandekerckhove, J., Verheyen, S., Tuerlinckx, F., (this issue). A crossed random effects diffusion model for speeded semantic categorization decisions. *Acta Psychologica*.

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology, 48*, 422–488.

Vinson, D., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods, 40*(1), 183–190.

Wallach, H. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd international conference on machine learning* (pp. 977–984). Pittsburgh, Pennsylvania, US.

Wang, C., Blei, D., & Heckerman, D. (2008). Continuous time dynamic topic models. In *Uncertainty in artificial intelligence (UAI).*

Zeigenfuse, M. D., & Lee, M. D. (this issue). Psychological contaminants as missing data: A latent assignment approach. *Acta Psychologica*.

Zeigenfuse, M. D., & Lee, M. D. (2008). Finding feature representations of stimuli: Combining feature generation and similarity judgment tasks. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (pp. 1825–1830). Austin, TX: Cognitive Science Society.

Please cite this article in press as: Steyvers, M. Combining feature norms and text data with topic models. *Acta Psychologica* (2009), doi:10.1016/j.actpsy.2009.10.010