TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 3 (2011) 18–47

ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2010.01097.x

Combining Background Knowledge and Learned Topics

Mark Steyvers,[a] Padhraic Smyth,[b] Chaitanya Chemuduganta,[b]

[a]*Department of Cognitive Sciences, University of California, Irvine*
[b]*Department of Computer Science, University of California, Irvine*

Received 12 November 2008; received in revised form 19 June 2009; accepted 08 September 2009

Abstract

Statistical topic models provide a general data-driven framework for automated discovery of high-level knowledge from large collections of text documents. Although topic models can potentially discover a broad range of themes in a data set, the interpretability of the learned topics is not always ideal. Human-defined concepts, however, tend to be semantically richer due to careful selection of words that define the concepts, but they may not span the themes in a data set exhaustively. In this study, we review a new probabilistic framework for combining a hierarchy of human-defined semantic concepts with a statistical topic model to seek the best of both worlds. Results indicate that this combination leads to systematic improvements in generalization performance as well as enabling new techniques for inferring and visualizing the content of a document.

Keywords: Topic model; Concept-topic model; Hierarchical concept-topic model; Concepts; Background knowledge; Human-defined knowledge; Data-driven learning; Bayesian models

1. Introduction

Many recent computational approaches to semantic cognition and statistical natural language processing operate on a purely data-driven basis. These models can extract useful information merely on the basis of statistical information contained in large text collections. From a machine-learning perspective, such models are attractive because they allow for a rapid analysis and understanding of new collections of text without significant human coding or annotation effort (e.g., Newman, Chemudugunta, Smyth, & Steyvers, 2006). From a cognitive science perspective, these models are attractive because they show that many findings related to semantic cognition can be explained by simple statistical learning processes.

Correspondence should be sent to Mark Steyvers, Department of Cognitive Sciences, University of California, 3151 Social Sciences Plaza, Irvine, CA 92697-5100. E-mail: mark.steyvers@uci.edu

Such learning processes can account for many empirical findings in areas such as language acquisition (Newport & Aslin, 2004; Newport, Hauser, Spaepen, & Aslin, 2004), multimodal language learning (Yu, Ballard, & Aslin, 2005), object perception (Fiser & Aslin, 2005), and eye movements (Najemnik & Geisler, 2005).

In this research, we start with the assumption that much of our semantic representations can be acquired from experience in the form of large text collections, given appropriate statistical learning machinery. However, we also assume that building in some structure and prior knowledge might be required to create suitable representations. It has been shown recently how data-driven learning approaches can be combined with structured representations such as hierarchies, graphs, trees, and rules to create powerful new learning models (Chater & Manning, 2006; Kemp & Tenenbaum, 2008). In our research, we show how structured background knowledge and statistical learning processes can be combined. The combination of prior knowledge and novel information gained from experience raises two broad theoretical questions. First, how can prior knowledge facilitate the acquisition of new knowledge? We will investigate the circumstances under which prior knowledge can significantly help in learning semantic representations. Second, how can new knowledge be used to make changes in our background knowledge? We will demonstrate how corpus-driven learning processes can be used to identify gaps in an existing semantic representation.

1.1. Data-driven learning approaches

There are a variety of unsupervised approaches for extracting semantic representations from large text collections that do not rely on background knowledge. In the context of a general ''bag-of-words'' framework, each document is represented by a vector that contains counts of the number of times each term (i.e., word or word combination) appears in the 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 (Kohonen et al., 2000; Lagus, Honkela, Kaski, & Kohonen, 1999) or linear projection methods such as latent semantic analysis (LSA; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). As a result of the dimensionality reduction, neighboring points in the semantic space often represent words or documents with similar contextual usages or meaning. These representations have been shown to model human knowledge in a variety of cognitive tasks (Landauer & Dumais, 1997) and educational assessment applications (Foltz, Gilliam, & Kendall, 2000). Other recent models in cognitive science have focused on alternative unsupervised methods to extract semantic representations at the sentence or document level (e.g., Dennis, 2004; Jones & Mewhort, 2007).

In a probabilistic framework, a variety of clustering techniques have been developed that characterize each document by a single latent cluster or topic (e.g., Cutting, Karger, Pedersen, & Tukey, 1992; McCallum, Nigam, & Ungar, 2000; Popescul, Ungar, Flake, Lawrence, & Giles, 2000). Through unsupervised learning, these clusters can be learned automatically and give broad information about the content of documents. The drawback of the one-to-one mapping between documents and clusters 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 et al., 2006).

A more flexible unsupervised framework, known as statistical topic modeling, allows each document to be represented by multiple topics (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 topic–document 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 author-topic models (Steyvers, Smyth, Rosen-Zvi, & Griffiths, 2004), correlated topics (Blei & Lafferty, 2006), hierarchical topic models (Blei, Griffiths, Jordan, & Tenenbaum, 2003; Li, Blei, & McCallum, 2007; Teh, Jordan, Beal, & Blei, 2006), time-dependent topics (Wang, Blei, & Heckerman, 2008) and models that combine topics and syntax (Griffiths, Steyvers, Blei, & Tenenbaum, 2005), as well as image features and text (Blei & Jordan, 2003). Topic models have also been useful as cognitive models to explain human associations, gist extraction, and memory errors (Griffiths et al., 2007).

One of the drawbacks of a purely data-driven learning process, such as topic modeling, is that the resulting representations can require some effort to interpret. As an illustrative example of the information learned by topic models, Fig. 1 (top row) shows five examples of topics that were derived from the TASA corpus, a collection of over 37,000 text passages from educational materials (e.g., language and arts, social studies, health, sciences) collected by Touchstone Applied Science Associates (TASA; see Landauer et al., 1998). The figure shows the 15 words that have the highest probability under each topic. Each number corresponds to the probability that a word is generated conditioned on a learned topic. It is often easier to interpret topics relative to other representations such as document clusters in cluster models or latent dimensions in latent semantic analysis (e.g., Newman et al., 2006). The words in the topics in the top row of Fig. 1 appear to relate to colors, gases, and the atmosphere, American presidents, European countries in World War II, and Japan and World War II, respectively. However, because topics are defined by probability distributions over words and have no simple names or definitions that can explain their content, an interpretation of the content of a topic often requires a subjective analysis of the connections between the high-probability words in a topic. This subjective process can lead to different outcomes depending on which individual is doing the analysis. Some progress has been made to automate the labeling of topics (Mei, Shen, & Zhai, 2007), but it remains to be seen how easily accessible such statistical representations are to human users.

Even with these techniques, topic interpretability remains an issue when faced with small noisy data sets. Data-driven learning models require large amounts of data in order to obtain accurate and useful representations, and such data might not always be available. In addition, because the model tunes itself to the dominant semantic themes in a corpus, it might not accurately represent outlier documents. Although such models might be able to tell that

word	prob.
red	0.202
blue	0.099
green	0.096
yellow	0.073
white	0.048
color	0.048
bright	0.030
colors	0.029
orange	0.027
brown	0.027
pink	0.017
look	0.017
black	0.016
purple	0.015
cross	0.011

word	prob.
oxygen	0.136
carbon	0.097
dioxide	0.050
air	0.046
ramona	0.037
gas	0.036
nitrogen	0.030
gases	0.026
atmosphere	0.020
hydrogen	0.020
water	0.016
respiraion	0.014
process	0.014
beezus	0.012
breathe	0.011

word	prob.
president	0.129
roosevelt	0.032
congress	0.030
johnson	0.026
office	0.021
wilson	0.021
nixon	0.020
reagan	0.018
kennedy	0.018
carter	0.017
presidents	0.012
administration	0.012
presidential	0.011
white	0.011
budget	0.010

word	prob.
france	0.071
french	0.069
europe	0.051
germany	0.043
german	0.041
countries	0.030
britain	0.024
italy	0.019
western	0.019
european	0.019
british	0.016
war	0.015
germans	0.013
country	0.012
nations	0.012

word	prob.
war	0.201
japanese	0.035
japan	0.035
II	0.035
american	0.030
peace	0.029
civil	0.019
end	0.016
wars	0.014
treaty	0.013
fought	0.012
fighting	0.012
military	0.012
ended	0.011
forces	0.011

COMMON COLORS
red
green
blue
brown
yellow
orange
pink
purple
reddish
yellowish
greenish
brownish
bluish
redness
pinkish

CHEMICAL ELEMENTS
oxygen
gold
iron
lead
carbon
hydrogen
copper
mercury
nitrogen
sodium
tin
aluminum
sulfur
calcium
uranium

LEADERS OF NATIONAL AND REGIONAL GOVERNMENTS
president
governor
presidential
ruler
presidency
dynasty
sovereign
chancellor
premier
governorship
regent
dynastic
gubernatorial
vp
mp

COUNTRY NAMES
america
england
france
china
mexico
india
spain
canada
japan
germany
egypt
russia
italy
alaska
australia

WAR
war
peace
pacific
campaign
peaceful
hostile
warfare
wartime
peacefully
tactics
concord
crusade
warlike
warring
peacetime

Fig. 1. Examples of five topics (out of 300) extracted from the TASA corpus (top row). The closest corresponding concepts from the CALD knowledge base (bottom row). The most likely words in each topic along with the matching word in the concept are highlighted. The column ''prob'' for each topic refers to the probability of each word in that topic.

a document falls outside the scope of representative semantic themes, it is difficult to identify the content that is covered by such documents. Therefore, in the absence of a large repository of relevant background documents to build topic models, it can be difficult to get interpretable and effective representations.

1.2. Human-defined semantic representations

An entirely different approach to constructing semantic representations is to rely on human knowledge and judgment. Considerable effort has gone into developing human-defined knowledge databases that characterize commonsense and lexical knowledge in humans. Such knowledge is created by trained experts in projects such as Cyc (Lenat & Guha, 1989; Panton et al., 2006), WordNet (Fellbaum, 1998; Miller, 1990), and Roget's thesaurus (Roget, 1911) or by untrained volunteers in projects such as ConceptNet (Havasi, Speer, & Alonso, 2007). Similarly, in cognitive science, many behavioral experiments have elicited detailed knowledge from many college students about semantic associations (Nelson, McEvoy, & Schreiber, 1998) and concepts and features (McRae, Cree, Seidenberg, & McNorgan, 2005; Ruts et al., 2004). Such human-defined representations can serve as proxies for mental representations.

To highlight the difference between learned topics and human-defined knowledge, we will show some examples of human-defined concepts that were created by lexicographers as part of the Cambridge Advanced Learner's Dictionary (CALD). In contrast to other taxonomies such as WordNet (Fellbaum, 1998; Miller, 1995), CALD groups words primarily according to semantic topics with the topics hierarchically organized. In addition, CALD provides names for each concept that are helpful for visualization purposes. CALD consists of 2,183 semantic concepts with each concept consisting of a set of words and a name that describes the concept. Fig. 1 (bottom row) shows an example of CALD concepts that resemble the meaning of the learned topics in the top row of Fig. 1. Each concept is illustrated with the name of the concept (shown on top) and a subset of 15 words that are part of the concept. Apart from alphabetical order, there is no natural way to order words within a concept. To better summarize the sometimes large number of words in each concept, we ordered the words by word frequency in the TASA corpus.

A comparison between learned topics and human-defined CALD concepts in Fig. 1 reveals some interesting differences: The words in each topic are associated with probabilities that indicate how likely each word is to be found in a context of that topic, which is quite useful to get a fine-grained indication about the relevance of a word to a topic. In contrast, CALD concepts do not provide any information about the prominence, frequency, or representativeness of the words in each concept—either a word is present or it is absent in a concept.

A clear advantage of concepts is that they are often more interpretable than learned topics by virtue of having a name (or small number of words) that describes the concept, providing concepts more precise coverage compared to topics. For example, the first topic on colors includes words that are not color words (e.g., *bright* and *look*), whereas a color concept will restrict itself to just color words. Concepts can also have broader coverage relative to topics because all words are considered as candidates and not just words occurring in a particular corpus. For example, the concept on CHEMICAL ELEMENTS lists all chemical elements (as known by the lexicographer), whereas a learned topic might focus more on the high-frequency chemical elements. Also, a learned topic could omit certain elements altogether because they did not occur in the corpus.

Although there are many advantages of human-defined knowledge databases, a major drawback is that they require extensive manual involvement and are time consuming to build and update given new emerging information. For some applications such as analyzing and summarizing text collections, no suitable knowledge database might even be available that has a suitable coverage of the domain. In contrast, data-driven topics can be tuned to themes in a corpus and can easily discover and summarize the dominant semantic themes for a wide variety of text collections.

1.3. Combining human-defined knowledge and data-driven learning approaches

Clearly, representations based on either a purely data-driven approach or human-defined knowledge have limitations. In this article, we will review some of our recent work that combines human-defined concepts with statistical data-driven approaches to learning

semantic representations (Chemudugunta, Holloway, Smyth, & Steyvers, 2008; Chemudugunta, Smyth, & Steyvers, 2008a, 2008b). The objective is to combine these approaches with the goal of taking advantage of the best features of both approaches. When there are few documents to learn from, these hybrid models are primarily driven by human-defined concepts. When trained on large document collections, data-driven topics can fill in gaps in the human-defined concepts. From a machine-learning perspective, automatically identifying such gaps can lead to a variety of useful applications where we update existing representations without requiring extensive human effort in discovering new emerging themes. From a cognitive science perspective, the hybrid model leads to novel ways of thinking about semantic representations. Instead of assuming that such representations are purely the result of data-driven learning processes, they might be a combination of preexisting knowledge and new knowledge extracted from a collection of text. We make no theoretical claims about the source of the prior knowledge. Although it is likely that such prior knowledge is itself acquired by experience, we do not attempt to explain how this is learned from experience.

The plan for the rest of the study is as follows. In Section 2, we review the basic principles of topic models and then describe the concept–topic model that combines concepts and topics into a single probabilistic model. We also describe the hierarchical concept–topic model which takes advantage of known hierarchical structure among concepts. Section 3 describes the text corpus and concept data set that we used to conduct our experiments. Section 4 describes a series of experiments that evaluate the predictive performance of a number of different models, showing for example that prior knowledge of concept words and concept relations can lead to better topic-based language models. In Section 5, we discuss a number of examples that illustrate how documents can be tagged at the word level with human-defined concepts. In Section 6, we discuss the type of information that is learned by topics but not captured by concepts. In Section 7, we show how the concept–topic model can automatically find appropriate concepts for novel words. In the final sections, we conclude the study with a brief discussion of related research, future directions, and final comments.

2. Concept–topic models

A clear advantage of an unsupervised learning approach such as topic modeling is that the model can be tuned to the themes of the particular document collection it is trained on. In addition, the probabilistic model that underlies the topic model allows one to automatically tag each word in a document with the topic most likely to have generated it. On the contrary, human-defined concepts such as the CALD knowledge base have much broader coverage of English words and include useful names of concepts that clarify the set of words that could be included in the concept, and aid in interpretability.

In this section, we will describe concept–topic and hierarchical concept–topic models that combine data-driven topics and human-defined concepts (Chemudugunta et al., 2008b, 2008c). We begin with a brief review of topic models.

2.1. Topic model

The topic model (or latent Dirichlet allocation [LDA] model; Blei et al., 2003) 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, $\phi^{(t)} = [\varphi_1^{(t)}, ..., \varphi_V^{(t)}]$, where $\varphi_w^{(t)} = p(w|t)$ and $1 \le w \le V$. 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 a document d is represented as a multinomial distribution over those T topics, $\theta^{(d)} = [\theta_1^{(d)}, ..., \theta_T^{(d)}]$, where $\theta_t^{(d)} = p(t|d)$ and $1 \le t \le T$. The variables φ and θ indicate which words are important for which topic and which topics are important for a particular document, respectively.

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 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:

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

In the LDA model, Dirichlet priors are placed on both φ and θ, to smooth the word–topic and topic–document distributions (for a description of Dirichlet priors, see Steyvers & Griffiths, 2007; Gelman, Carlin, Stern, & Rubin, 2003). In many applications, a symmetric Dirichlet density with single hyperparameters α and β are used for θ and φ, respectively. For all the topic models in this research, we will use a symmetric Dirichlet prior for φ using a single hyperparameter β. For the topic–document distributions θ, we will use an asymmetric Dirichlet prior θ, with a vector $\boldsymbol{\alpha}$ containing hyperparameter values for every topic (and concept for concept–topic models). An asymmetric prior is useful when some concepts (or 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 to express rare or frequent topics (or concepts).

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

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

This generative process can be summarized by the graphical model shown in Fig. 2A. In the 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 word tokens 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 distributions over word types 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 z of individual words to topics. These distributions can be learned in a completely unsupervised

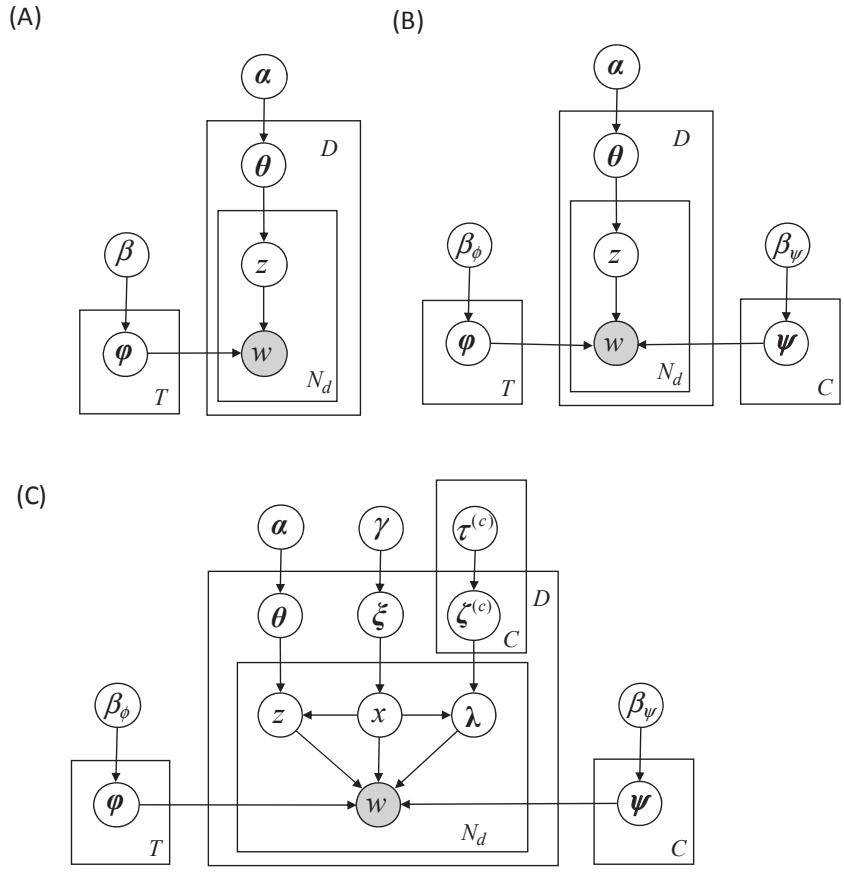

Fig. 2. Graphical models for the topic model (A), the concept–topic model (B), and the hierarchical concept–topic model (C).

manner without any prior knowledge about topics or which topics are covered by which documents. One efficient technique for obtaining estimates of these distributions is through collapsed Gibbs sampling (Griffiths & Steyvers, 2004). Steyvers and Griffiths (2007) present a tutorial introduction to topic models that discusses collapsed Gibbs sampling. The main idea of collapsed Gibbs sampling is that inference is performed only on z, the assignments of word tokens to topics. The remaining latent variables θ and φ are integrated out (''collapsed''). 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 are then used to estimate the word–topic and topic–document distributions. The vector $\boldsymbol{\alpha}$ that contains the hyperparameter values for every topic (and concept for concept–topic models, see below) is updated using a process involving fixed-point update equations (Minka, 2000; Wallach, 2006). See Appendix A of Chemudugunta et al. (2008b) for more details.

To summarize, the topic model provides several pieces of information that are useful for understanding documents. The topic–document distributions indicate the important topics in each document. The word–topic distributions indicate which words are important for which topic (e.g., the top row of Fig. 1 shows some example word–topic distributions estimated for the TASA corpus). Finally, the probabilistic assignments z_i of word tokens to topics are useful for tagging purposes, providing information about the role each word is playing in a specific document context and helping to disambiguate multiple meanings of a word (e.g., Griffiths et al., 2007).

2.2. Concept–topic model

The concept–topic model is a simple extension to the topic model where we add C concepts to the T topics of the topic model resulting in an effective set of $T + C$ word distributions for each document. We assume that each of the C concepts (such as the CALD concepts in Fig. 1) are represented as a set of words. Therefore, these human-defined concepts only give us a membership function over words—either a word is a member of the concept or it is not. One straightforward way to incorporate concepts into the topic modeling framework is to convert them to probability distributions over their associated word sets. In the concept–topic model, we will treat each concept c as a multinomial distribution $\psi^{(c)} = [\psi_1^{(c)}, ..., \psi_V^{(c)}]$, where $\psi_w^{(c)} = p(w|c)$ and $1 \leq w \leq V$. Importantly, each word type that is not part of the concept will have zero probability, that is, $\psi_w^{(c)} = 0$ for $w \notin c$. Of course, there are no direct observations available about the probabilities of word types within a concept, but we can use a model similar to the topic model to estimate these probabilities from corpus data. Therefore, the concept–topic model is simply an extension of the topic model where we have a number of learned topics as well as constrained topics where nonzero probability can only be given to words in human-defined concepts.

In the concept–topic model, the conditional probability of the ith word token w_i in a document d is

$$p(w_i|d) = \sum_{t=1}^{T} p(w_i|z_i = t)p(z_i = t|d) + \sum_{t=T+1}^{C+T} p(w_i|z_i = t)p(z_i = t|d), \quad (2)$$

where the indices $1 \le t \le T$ refer to all topics and indices $T + 1 \le t \le T + C$ refer to all concepts. In this generative process, an index z_i is sampled from the distribution over topics and concepts for the particular document. If $z_i \le T$, a word token is sampled from topic z_i, and if $T + 1 \le z_i \le T + C$, a word token is sampled from concept $z_i - T$ among word types associated with the concept. The topic model can be viewed as a special case of the concept–topic model when there are no concepts present, that is, when $C = 0$. At the other extreme of this model where $T = 0$, the model relies entirely on predefined concepts.

To specify the complete generative model, let $\phi^{(t)} = [\varphi_1^{(t)}, ..., \varphi_V^{(t)}]$, where $\varphi_w^{(t)} = p(w|t)$ and $1 \le w \le V$, refer to the multinomial distribution over word types for topic t when $1 \le t \le T$, and let $\Psi^{(c)} = [\psi_1^{(c)}, ..., \psi_V^{(c)}]$, where $\psi_w^{(c)} = p(w|c)$ and $1 \le w \le V$ refer to the multinomial distribution over word types for concept $c = t–T$ when $T + 1 \le t \le T + C$. As with the topic model, we place Dirichlet priors on the multinomial variables θ, φ, and ψ, with corresponding hyperparameters $\boldsymbol{\alpha}$, β_ϕ, and β_ψ.

The complete generative process can be described as follows:

1. For each topic $t \in \{1, ..., T\}$, select a word distribution $\phi^{(t)} \sim \text{Dirichlet}(\beta_\phi)$
2. For each concept $c \in \{1, ..., C\}$, select a word distribution $\psi^{(c)} \sim \text{Dirichlet}(\beta_\psi)$
3. For each document $d \in \{1, ..., D\}$
   (a) Select a distribution over topics and concepts $\theta^{(d)} \sim \text{Dirichlet}(\alpha)$
   (b) For each word position i in document d
      (i) Select a component $z_i \sim \text{Discrete}(\theta^{(d)})$
      (ii) If $z_i \le T$, generate a word token from topic z_i, $w_i \sim \text{Discrete}(\phi^{(z_i)})$; otherwise, generate a word token from concept $c_i = z_i - T$, $w_i \sim \text{Discrete}(\psi^{(c_i)})$

Note that in Step 2, the sampling of words for a concept is constrained to only the words that are members of the human-defined concept. Fig. 2B shows the corresponding graphical model. All the latent variables in the model can be inferred through collapsed Gibbs sampling in a similar manner to the topic model (see Chemudugunta et al., 2008b for details).

We note that even though we are partially relying on humans to define the word–concept memberships, we still apply purely unsupervised algorithms to estimate the latent variables in the model. This is in contrast to a supervised learning approach where the human-defined knowledge is used as a target for prediction. Here, the human-defined knowledge is only used as a constraint on the probability distributions that can be learned for each concept.

We also note that the concept–topic model is not the only way to incorporate semantic concepts. For example, we could use the concept–word associations to build informative priors for the topic model and then allow the inference algorithm to learn word probabilities for all words (for each concept), given the prior and the data. We chose the restricted vocabulary approach to exploit the sparsity in the concept–word associations (topics are distributions over all the words in the vocabulary but concepts are restricted to just their sets of associated words, which are much smaller than the full vocabulary). This sparsity at the

word level allows us to easily perform inference with tens of thousands of concepts on large document collections.

A general motivation for the concept–topic approach is that there might be topics present in a corpus that are not represented in the concept set (but that can be learned). Similarly, there may be concepts that are either missing from the text corpus or are rare enough that they are not found in the data-driven topics of the topic model. The marriage of concepts and topics provides a simple way to augment concepts with topics and has the flexibility to mix and match topics and concepts to describe a document.

2.3. Hierarchical concept–topic model

Although the concept–topic model provides a simple way to combine concepts and topics, it does not take into account any hierarchical structure the concepts might have. For example, CALD concepts are arranged in a hierarchy that starts with the concept EVERYTHING which splits into 17 concepts at the second level (e.g., SCIENCE, SOCIETY, GENERAL/ABSTRACT, COMMUNICATION). The hierarchy has up to seven levels with each level specifying more specific concepts.

In this section, we describe a hierarchical concept–topic model that incorporates hierarchical structure of a concept set. Similar to the concept–topic model described in the previous section, there are T topics and C concepts. However, as opposed to the flat organization of the concepts in the concept–topic model, we now utilize the hierarchical organization of concepts when sampling words from concepts. Before we formally describe the model, we illustrate the basic idea in Fig. 3. Each topic and concept is associated with a ''bag of words'' that represents a multinomial distribution over word types. In the generative process, word tokens can be generated from the concept part of the model by sampling a path from the root of the concept tree to some distribution over word types associated with the concept (left box in Fig. 3). Alternatively, word tokens can be generated through the topic part of the model (right box). The dashed and dotted lines show examples of two word tokens sampled through the hierarchical concept part of the model and the topic part of the model, respectively. For the first word token, the option ''topic'' is sampled at the root node, Topic 1 is then sampled, and then a word token is sampled from the multinomial over words associated with Topic 1. For the second word token, the option ''concept'' is sampled at the root node, then the option SCIENCE is sampled as a child of the concept EVERYTHING, the word distribution for SCIENCE is then selected, and a word from this distribution is sampled. Each transition in the hierarchical part of the model has an associated probability and the transition probabilities are document dependent—some paths are more likely in context of some documents. For example, in physics and chemistry documents, one might expect all transitions toward the SCIENCE concept to be elevated but differentiated between the transitions toward the PHYSICS and CHEMISTRY concepts.

To preview what information is learned by the model, we need to distinguish between variables learned at the word, document, and corpus levels. At the word level, the model learns the assignments of topics or concepts to word tokens. These assignments can be directly used for tagging purposes and word–sense disambiguation. At the document level,

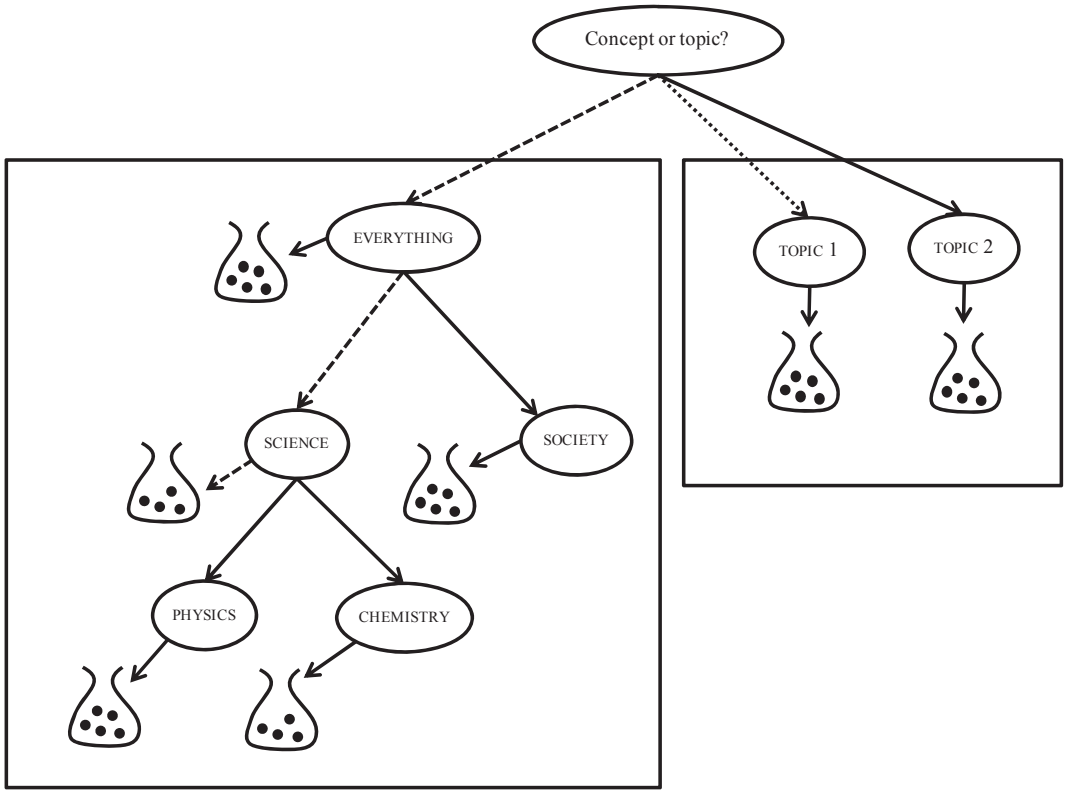

Fig. 3. An illustration of the hierarchical concept–topic model.

the model learns both topic probabilities and concept–transition probabilities in the concept tree. The latter information is useful because it allows a hierarchical representation of document content. At the document level, the model also learns the switch probability that a word is generated through the topic or concept route. The adaptive nature of the switch probability allows the model to flexibly adapt to different documents. Documents that contain material that has poor concept coverage will have a high probability of switching to the topic route. At the corpus level, the model learns the probabilities of the word–topic and word–concept distributions. The word–topic distributions are useful to learn which semantic themes beyond those covered in the concepts are needed to explain the content of the whole document collections. The word–concept distributions are useful to learn which words are important for each concept. Finally, at the corpus level, the model also learns the hyperparameters for each transition in the concept tree. The learned hyperparameters allow the model to make certain paths more prominent across all documents. For example, if a document collection includes many documents on science, the path toward the SCIENCE concept could be made more likely (a priori).

Our approach is related to the hierarchical pachinko allocation model 2 (HPAM 2) as described by Mimno, Li, and McCallum (2007). In the HPAM 2 model, topics are arranged

in a three-level hierarchy with root, super-topics, and subtopics at Levels 1, 2, and 3, respectively, and words are generated by traversing the topic hierarchy and exiting at a specific level and node. In our model, we use a similar mechanism for word generation via the concept route. There is additional machinery in our model to incorporate the data-driven topics (in addition to the hierarchy of concepts) and a switching mechanism to choose the word generation process via the concept route or the topic route.

To give a formal description of model, for each document d, we introduce a ''switch'' distribution $p(x|d)$ that determines if a word should be generated via the topic route or the concept route. Every word token w_i in the corpus is associated with a binary switch variable x_i. If $x_i = 0$, the previously described standard topic mechanism is used to generate the word. That is, we first select a topic t from a document-specific mixture of topics $\theta^{(d)}$ and generate a word token from the word distribution associated with topic t. If $x_i = 1$, we generate the word token from one of the C concepts in the concept tree. To do that, we associate with each concept node c in the concept tree a document-specific multinomial distribution with dimensionality equal to $N_c + 1$, where N_c is the number of children of the concept node c. This distribution allows us to traverse the concept tree and exit at any of the C nodes in the tree—given that we are at a concept node c, there are N_c child concepts to choose from and an additional option to choose an ''exit'' child to exit the concept tree at concept node c. We start our walk through the concept tree at the root node and select a child node from one of its children. We repeat this process until we reach an exit node and a word token is generated from the parent of the exit node. Note that for a concept tree with C nodes, there are exactly C distinct ways to select a path and exit the tree, as there is only one parent for each concept node, and thus, one path to each of the C concepts.

In the hierarchical concept–topic model, a document is represented as a weighted combination of mixtures of T topics and C paths through the concept tree and the conditional probability of the ith word token in document d is given by

$$\begin{aligned} p(w_i|d) &= p(x_i = 0|d) \sum_{t=1}^{T} p(w_i|z_i = t)p(z_i = t|d) \\ &+ p(x_i = 1|d) \sum_{c=T+1}^{T+C} p(w_i|z_i = c)[p(\text{exit}|c, d)p(c|\text{parent}(c), d)\cdots p(\text{root}|d)] \end{aligned} \tag{3}$$

The sequential process to generate a document collection with D documents under the hierarchical concept–topic model is as follows:

1. For each topic $t \in \{1, ..., T\}$, select a word distribution $\phi^{(t)} \sim \text{Dirichlet}(\beta_\phi)$
2. For each concept $c \in \{1, ..., C\}$, select a word distribution $\psi^{(c)} \sim \text{Dirichlet}(\beta_\psi)$
3. For each document $d \in \{1, ..., D\}$
   (a) Select a switch distribution $\xi^{(d)} \sim \text{Beta}(\gamma)$
   (b) Select a distribution over topics $\theta^{(d)} \sim \text{Dirichlet}(\alpha)$
   (c) For each concept $c \in \{1, ..., C\}$
      i. Select a distribution over children of c, $\zeta^{(cd)} \sim \text{Dirichlet}(\tau^{(c)})$

(d) For each word position i in document d
(i) Select a binary switch $x_i \sim \text{Bernoulli}\left(\xi^{(d)}\right)$
(ii) If $x_i = 0$
(A) Select a topic $z_i \sim \text{Discrete}\left(\theta^{(d)}\right)$
(B) Generate a word from topic z_i, $w_i \sim \text{Discrete}\left(\phi^{(z_i)}\right)$
(iii) Otherwise, create a path starting at the root concept node, $\lambda_1 = 1$
(A) Select a child node $\lambda_j, \lambda_{j+1} \sim \text{Discrete}(\zeta^{(\lambda_j d)})$, and increment j. Repeat until $\lambda_{j+1}$is an exit node
(B) Generate a word from concept $c_i = \lambda_j$, $w_i \sim \text{Discrete}\left(\psi^{(c_i)}\right)$. Set $z_i = c_i + T$

where $\varphi^{(t)}$, $\psi^{(c)}$, β_φ, and β_ψ are analogous to the corresponding symbols in the concept–topic model described in the previous section. The variable $\xi^{(d)}$, where $\xi^{(d)} = p(x|d)$, represents the switch distribution and $\theta^{(d)}$, where $\theta^{(d)} = p(t|d)$ represents the distribution over topics for document d. The variable $\zeta^{(cd)}$represents the multinomial distribution over children of concept node c for document d (this has dimensionality $N_c + 1$ to account for the additional ''exit'' child). The hyperparameters γ, $\boldsymbol{\alpha}$, and $\boldsymbol{\tau}^{(c)}$ are the parameters of the priors on $\xi^{(d)}$, $\theta^{(d)}$, and $\zeta^{(cd)}$, respectively. Note that $\boldsymbol{\alpha}$, as in the previous topic and concept–topic models, is a vector with hyperparameter values for each topic. Similarly, $\boldsymbol{\tau}^{(c)}$is a vector of hyperparameters values, to allow for different a priori probabilities of traversing the concept-tree. This allows the model to tune itself to different corpora and make it more likely to sample a path toward the SCIENCE concept in a corpus of scientific documents. Fig. 2C shows the corresponding graphical model. The generative process above is quite flexible and can handle any directed-acyclic concept graph (for any nontree, there would be more than one way of reaching each concept, leading to increased complexity in the inference process). The model cannot, however, handle cycles in the concept structure as the walk of the concept graph starting at the root node is not guaranteed to terminate at an exit node.

In the hierarchical concept–topic model, the only observed information is the set of words in each document, the word–concept memberships, and the tree structure of the concepts. All remaining variables are latent and are inferred through a collapsed Gibbs sampling procedure. Details about this procedure are described by Chemudugunta et al. (2008b).

3. Text and concept data

The experiments for all our simulations are based on the TASA corpus (Landauer & Dumais, 1997) consisting of D = 37,651 documents with passages excerpted from educational texts used in curricula from the first year of school to the first year of college. The documents are divided into nine different educational genres. We focus here on a subset of TASA documents classified as *science*, consisting of D = 5,356 documents. As mentioned previously, CALD consists of 2,183 semantic concepts. CALD groups words primarily according to semantic concepts with the concepts hierarchically organized. The hierarchy starts with the concept EVERYTHING which splits into 17 concepts at the second level (e.g.,

SCIENCE, SOCIETY, GENERAL/ABSTRACT, COMMUNICATION). The hierarchy has up to seven levels, with each interior node splitting into a median of seven children nodes. The concepts vary in the number of the words with a median of 54 word types and a maximum of 3,074. Each word can be a member of multiple concepts, especially if the word has multiple senses. We created two vocabularies. One is a $W = 21{,}072$ word vocabulary based on the *intersection* between the vocabularies of TASA and CALD. We also created a vocabulary of $W = 142{,}010$ words based on the *union* of TASA and CALD vocabularies. For both vocabularies, all stop words and infrequent words were removed.

4. Tagging documents

One application of concept models is to tag unlabeled documents with human-defined concepts. The tagging process involves assigning likely concepts to each word in a document, depending on the context of the document. The document content can then be summarized by the probability distribution over concepts that reveal the dominant semantic themes. Because the concept models assign concepts at the word level, the results can be aggregated in many ways, allowing for document summaries at multiple levels of granularity. For example, tagging can be performed on snippets of text, individual sections of a document, whole documents, or even collections of documents. For all of our tagging examples, we used the intersection vocabulary (the results are qualitatively similar using the union vocabulary).

4.1. Tagging with the concept–topic model

As an illustration of how the model can be used to quickly summarize a document, Fig. 4 shows the CALD concept assignments to individual words in a TASA document. We used the concept–topic model with concepts only ($T = 0$). The four most likely concepts are listed for this document. For each concept, the estimated probability distribution over words is shown next to the concept (note that these estimates are over the whole corpus and are not document specific). For example, for the concept of CHEMICAL ELEMENTS, the word *oxygen* is more likely than the word *chlorine*. The probability of words in concepts is not just influenced by number of tokens across the whole corpus but also by the number of concepts that contain the word type and the relative probability between concepts in each document. The model has estimated that in the conceptual context of CHEMICAL ELEMENTS, the word *oxygen* is more likely than the word *chlorine*. This conditional salience is useful for evaluating the relative importance of words to specific concepts, going beyond the logical set definitions provided by the human lexicographers who developed the concepts.

In the document, words assigned to the four most likely concepts are tagged with letters a–d (and color coded if viewing in color). The words assigned to any other concept are tagged with ''o'' and words outside the vocabulary are not tagged. In the concept–topic model, the distributions over concepts within a document are highly skewed such that the probability mass is distributed over only a small number of concepts. In the example

tag	P(c|d)	Concept	P(w|c)
a	0.1702	PHYSICS	electrons (0.2767) electron (0.1367) radiation (0.0899) protons (0.0723) ions (0.0532) radioactive (0.0476) proton (0.0282)
b	0.1325	CHEMICAL ELEMENTS	oxygen (0.3023) hydrogen (0.1871) carbon (0.0710) nitrogen (0.0670) sodium (0.0562) sulfur (0.0414) chlorine (0.0398)
c	0.0959	ATOMS, MOLECULES, AND SUB-ATOMIC PARTICLES	atoms (0.3009) molecules (0.2965) atom (0.2291) molecule (0.1085) ions (0.0262) isotopes (0.0135) ion (0.0105) isotope (0.0069)
d	0.0924	ELECTRICITY AND ELECTRONICS	electricity (0.2464) electric (0.2291) electrical (0.1082) current (0.0882) flow (0.0448) magnetism (0.0329)
o	0.5091	OTHER	

The hydrogen[b] ions[a] immediately[o] attach[o] themselves to water[o] molecules[c] to form[o] combinations[o] called[o] hydronium ions[a]. The chlorine[b] ions[a] also associate[o] with water[o] molecules[c] and become hydrated. Ordinarily[o], the positive[o] hydronium ions[a] and the negative[o] chlorine[b] ions[a] wander[o] about freely[o] in the solution[o] in all directions[o]. However, when the electrolytic cell[o] is connected[o] to a battery[o], the anode[d] becomes positively[o] charged[a] and the cathode[d] becomes negatively[o] charged[a]. The positively[o] charged[a] hydronium ions[a] are then attracted[o] toward the cathode[d] and the negatively[o] charged[a] chlorine[b] ions[a] are attracted[o] toward the anode[d]. The flow[d] of current[d] inside[o] the cell[o] therefore consists of positive[o] hydronium ions[a] flowing[d] in one direction[o] and negative[o] chlorine[b] ions[a] flowing[d] in the opposite[o] direction[o]. When the hydronium ions[a] reach[o] the cathode[d], which has an excess[o] of electrons[a], each takes[o] one electron[a] from it and thus neutralizes[o] the positively[o] charged[a] hydrogen[b] ion[a] attached[o] to it. The hydrogen[b] ions[a] thus become hydrogen[b] atoms[c] and are released[o] into the solution[o]. Here they pair[o] up to form[o] hydrogen[b] molecules[c] which gradually[o] come out of the solution[o] as bubbles[o] of hydrogen[b] gas[o]. When the chlorine[b] ions[a] reach[o] the anode[d], which has a shortage[o] of electrons[a], they give[o] up their extra[o] electrons[a] and become neutral[a] chlorine[b] atoms[c]. These pair[o] up to form[o] chlorine[b] molecules[c] which gradually[o] come out of the solution[o] as bubbles[o] of chlorine[b] gas[o]. The behavior[o] of hydrochloric acid[o] solution[o] is typical[o] of all electrolytes[o]. In general[o], when acids[o], bases[o], and salts[o] are dissolved[o] in water[o], many of their molecules[c] break[o] up into positively[o] and negatively[o] charged[a] ions[a] which are free[o] to move[o] in the solution[o].

Fig. 4. Illustrative example of tagging a document excerpt using the concept–topic model with concepts from CALD.

document, the four most likely concepts cover about 50% of all words in the document. Fig. 4 illustrates that the model correctly disambiguates between words that have several conceptual interpretations. For example, the word *charged* has many different meanings and appears in 20 CALD concepts. In the example document, this word is assigned to the PHYSICS concept, which is a reasonable interpretation in this document context (the word *charged* does not appear in the list of words associated with the concept PHYSICS because its probability falls below the threshold for visualization). Similarly, the ambiguous words *current* and *flow* are correctly assigned to the ELECTRICITY concept.

4.2. Tagging with the hierarchical concept–topic model

One of the advantages of the hierarchical concept–topic model is that the hierarchical relations between concepts can be used to enhance the visualization of documents. Fig. 5 shows the result of inferring the hierarchical concept mixture for an individual TASA document using CALD concept sets. For the hierarchy visualization, we selected the seven concepts with the highest probability and included all ancestors of these concepts when visualizing the tree (we selected seven concepts to tradeoff informativeness and complexity of the display). The ancestors that were not part of the top seven concepts are visualized with dashed ovals. The CALD subtree highlights the specific semantic themes of BIRTH, BREATHING, and STOPPING BREATHING along with the more general themes of SCIENCE and MEDICINE. This illustration shows that the model is able to give interpretable results for an individual document at multiple levels of granularity.

At a higher level of granularity, the hierarchical concept–topic model can also summarize sets of documents. Across documents, the model learns the hyperparameters associated with

(A)

The postnatal period of development lasts from birth until death and can be divided into a neonatal period, infancy, childhood, adolescence, adulthood, and senescence. The neonatal period, which extends from birth to the end of the first four weeks, begins very abruptly at birth. Physiological adjustments must be made quickly, because the newborn must suddenly do for itself those things that the mother body has been doing for it. Thus, the newborn must carry on respiration, obtain nutrients, digest nutrients, excrete wastes, regulate body temperature, and so forth. However, its most immediate need is to obtain oxygen and excrete carbon dioxide, so its first breath is critical. The first breath must be particularly forceful, because the newborn lungs are collapsed, and the airways are small and offer considerable resistance to air movement. Also, surface tension tends to hold the moist membranes of the lungs together. Fortunately, the lungs of a full term fetus secrete surfactant, which reduces surface tension, and after the first powerful breath begins to expand the lungs, breathing becomes easier. It is not clear whether the first breath is stimulated by one or several factors. Those that may be involved include an increasing level of carbon dioxide, a decreasing ph, low oxygen concentration, a drop in body temperature and mechanical stimulation that occurs during and after the birth process. Prior to birth, the fetus depends primarily on glucose and fatty acids obtained from the mother blood as energy sources

(B)

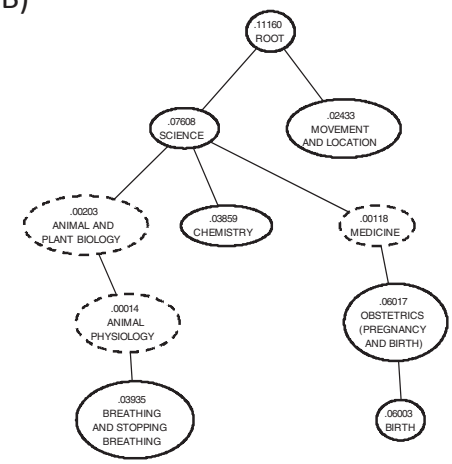

Fig. 5. Example of a single TASA document from the science genre (A). The seven-most probable concepts inferred by the hierarchical concept–topic model for this document using the CALD concepts (B). The dashed concepts are ancestor concepts of the top seven concepts that were included for visualization purposes.

the transitions from concept nodes to the children of concept nodes. These hyperparameters determine how likely it is (a priori) for the model to generate a document along the path to the PHYSICS and CHEMISTRY concepts (for example). Fig. 6 shows the 20 highest probability concepts for a random subset of 200 TASA documents from the science genre. For each concept, the name of the concept is shown in all caps. The visualization also includes the ancestor nodes (shown in dashed ovals) to complete the path to the root node. The numbers in Fig. 6 represent the marginal probability for the concept. The marginal probability is computed based on the product of probabilities along the path of reaching the node as well as the probability of exiting at the node, marginalized (averaged) across all documents:

$$p(c) \propto \sum_{d} [p(\text{exit}|c,d)p(c|\text{parent}(c),d)\cdots p(\text{root}|d)]. \tag{4}$$

Many of the most likely concepts as inferred by the model relate to specific science concepts (e.g., GEOGRAPHY, ASTRONOMY, CHEMISTRY, etc.). These concepts all also fall under the general SCIENCE concept, which is also one of the most likely concepts for this document collection. Therefore, the model is able to summarize the semantic themes in a set of documents at multiple levels of granularity.

In the original CALD concept set, each concept consists of a set of words and no knowledge is provided about the prominence, frequency, or representativeness of words within the concept. In the hierarchical concept–topic model, for each concept, a distribution over words is inferred that is tuned to the specific collection of documents. For example, for the concept ASTRONOMY, the word *planet* receives much higher probability than the word *Saturn* or *equinox*, all of which are members of the concept. These differences in word probabilities

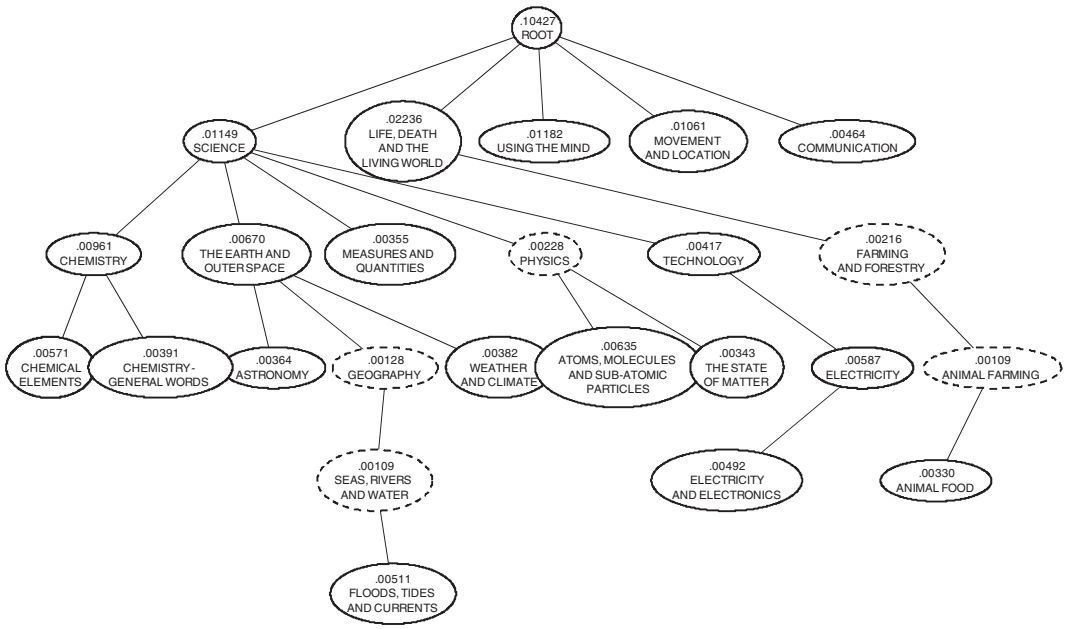

Fig. 6. Visualization of the marginal concept distributions from the hierarchical concept–topic model learned on science documents using CALD concepts. The 20 most likely concepts are shown, including the five ancestor nodes (shown in dashed ovals) needed to complete the path to the root node.

highlight the ability of the model to adapt to variations in word usage across document collections.

5. Generalization performance

In the previous section, the tagging illustrations provided a qualitative assessment of concept–topic models. To get a more quantitative evaluation, we assess the performance of the topic model, concept–topic model, and the hierarchical concept–topic model by evaluating their capability to explain new documents that the model has not been trained on. The idea is that models that are trained on documents of a certain genre should generalize to new documents from the same genre. One formal way to assess generalization performance is through *perplexity*. Perplexity is a quantitative measure for comparing language models (Brown, deSouza, Mercer, Della Pietra, & Lai, 1992) and is widely used to compare the predictive performance of topic models (e.g., Blei et al., 2003; Wallach et al., 2009). Although perplexity does not directly measure aspects of a model such as interpretability or coverage, it is nonetheless a useful general predictive metric for assessing the quality of a topic model.

Perplexity is equivalent to the inverse of the geometric mean of the likelihood of holdout data. The perplexity of a collection of test documents given the training set is defined as:

$$\mathrm{Perp}(\mathbf{w}_{\mathrm{test}}|D_{\mathrm{train}}) = \exp\left(-\frac{\sum_{d=1}^{D_{\mathrm{test}}} \log p(\mathbf{w}_d|D_{\mathrm{train}})}{\sum_{d=1}^{D_{\mathrm{test}}} N_d}\right) \quad (5)$$

where $\mathbf{w}_{\mathrm{test}}$ is the set of word tokens in the test documents, $\mathbf{w}_d$ is the set of word tokens in document d of the test set, D_{train} is the training set, and N_d is the number of word tokens in document d. Lower perplexity scores indicate that the model's predicted distribution of heldout data is closer to the true distribution.

The experiments in this section are again based on the TASA data set. We train the models on a random subset of 90% of documents classified as *science*, creating a training set of $D = 4{,}820$ documents. By training the models, we obtain estimates for the word–topic distributions, topic–document distributions, the assignments of word tokens to topics and concepts, as well as the hyperparameters on the topic–document distributions (for all models, asymmetric Dirichlet priors were used for the document-specific topic distributions). Note that in all reported simulations, we used the intersection vocabulary (the results are qualitatively similar using the union vocabulary).

We then evaluate generalization performance on the remaining documents in the science genre and also on a subset of documents classified as *social studies*. By testing on science and social studies documents, we evaluate the models' ability to generalize either within the same genre or between genres. For each test document, we used a random 50% of words of the document to estimate document-specific distributions and measure perplexity on the remaining 50% of words using the estimated distributions. More details about the perplexity computation are provided in Appendix B of Chemudugunta et al. (2008b).

5.1. Perplexity comparison across models

We compare the perplexity of the topic model (TM), concept–topic model (CTM), and the hierarchical concept–topic model (HCTM) trained on document sets from the science genre of the TASA collection and using concepts from CALD. Fig. 7A, B shows the perplexity of TM, CTM, and HCTM as a function of the number of data-driven topics T. Panel (a) shows the results when the model is trained and tested on science documents. Panel (b) shows the results when the model is trained on science documents and tested on social studies documents. The point $T = 0$ indicates that there are no topics used in the model. The results clearly indicate that incorporating concepts greatly improves the perplexity of the models (lower perplexity indicates better predictive performance). Both CTM and HCTM outperform TM, which does not rely on human-defined concepts. The results also show that human-defined concepts by themselves (i.e., the perplexity obtained when the number of learned topics $T = 0$) are not sufficient to get the best generalization performance—additional learned topics that are tuned to the specific content of the document collection are needed for optimal performance (around 100–300 learned topics). One important point to note is that the improved performance by the concept models is not due to the high number of word distributions $T + C$, compared with the topic model that utilizes only T topics. In fact, even with $T = 2{,}000$ topics, TM does not improve its perplexity and even shows signs of deterioration in quality.

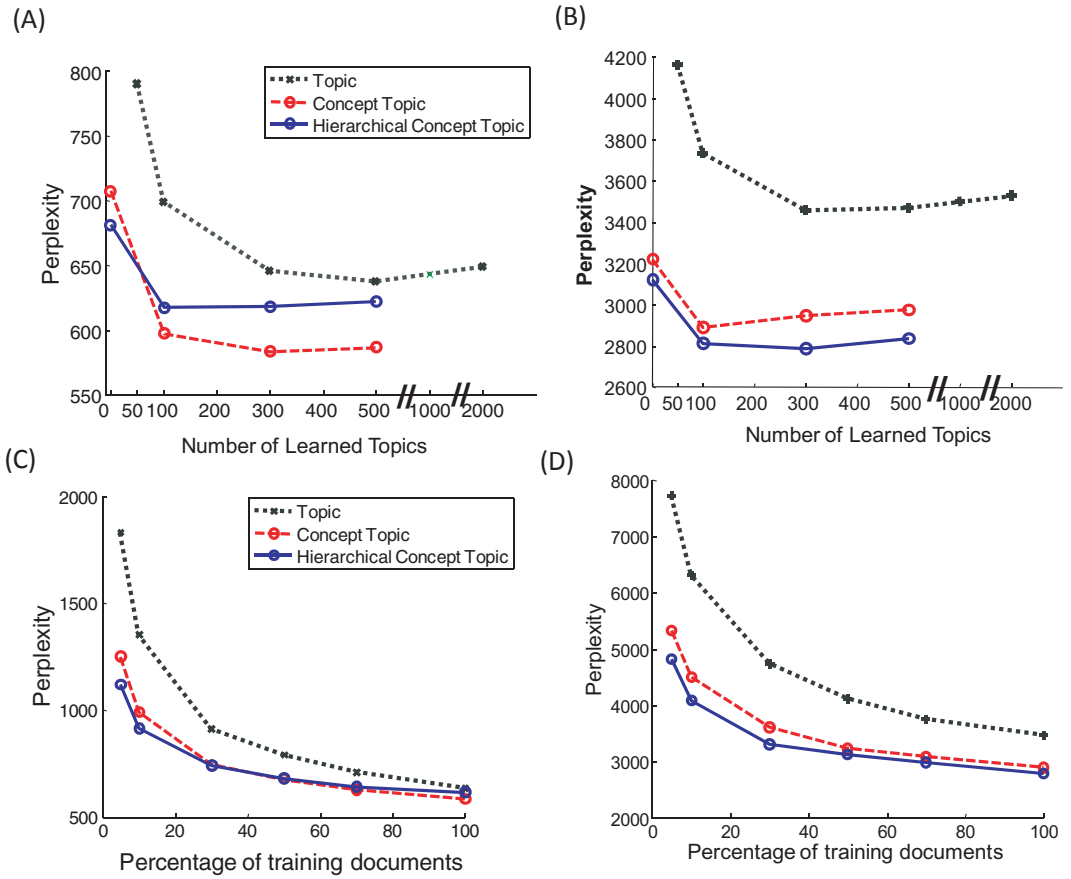

Fig. 7. Comparing perplexity for topic model, concept–topic model, and the hierarchical concept–topic model as a function of number of topics (A–B) and percentage of training documents (C–D). Panels (A) and (C) show the results when the model is trained and tested on documents from the science genre. Panels (B) and (D) show the results when the model is trained on documents from the science genre, but tested on documents from the social studies genre.

We next look at the effect of varying the amount of training data for all models. Fig. 7C shows the results when the model is trained and tested on science documents. Fig. 7D shows the results when the model is trained on science documents and tested on social studies documents. When there is very little training data (e.g., up to 500 documents), both concept–topic models significantly outperform the topic model. Because learned topics in TM are entirely data driven, there is not enough statistical information to build accurate representations on the basis of just a few hundred documents (in the extreme case where there is no training data available, topics of TM will just be uniform distributions and prediction will be at chance). In the regime of little training data, however, the concept models can leverage the human-defined concepts, providing a priori structure to the learning algorithm.

Of the two concept models, HCTM outperforms CTM when little training data are available (see Fig. 7A,B) or when the model generalizes to documents from a different genre (see Fig. 7B,D). HCTM and CTM rely on the same set of human-defined concepts, but HCTM also imposes hierarchical constraints on these concepts. For example, the concept model needs no documents to learn that physics and chemistry are related concepts since this knowledge is already built in. Therefore, if a document appears to be about physics, the model predicts with small probability that chemistry words can appear in the document. These a priori concept relations are clearly useful when little data are available but are relatively less beneficial with larger amounts of training data. In this scenario, the concepts and topics can be fine-tuned to the data and the difference in performance between flat concept–topic representations and hierarchical concept representations are less pronounced. When generalizing to new kinds of documents (e.g., when training on science documents and testing on social studies documents), the hierarchical concept–topic model outperforms the concept–topic model regardless of the amount of training data. In this case, the learned knowledge is less useful and the a priori structure in the hierarchical relations between concepts provides necessary constraints on the inference process.

6. Relation between learned topics and concepts

Both the concept–topic and hierarchical concept–topic models allow for a combination of concepts and learned topics. These learned topics are useful to identify different gaps in the existing concept sets and capture semantic themes beyond those covered in the concepts. Such learned topics will depend on the background corpus. To get a better understanding of the kind of information captured by these models, we applied the concept–topic model to the TASA documents in the *science* genre. We set the number of learned topics $T = 50$. In one simulation, we ran the model using the union vocabulary that combines words from the TASA corpus and CALD concepts. Importantly, the union vocabulary includes words that are not part of CALD. Because the model gives zero probability for such words under any concepts, these words have to be modeled by the learned topics. Fig. 8A shows examples of topics learned by the model. The words not covered by CALD are shown in bold. The learned topics clearly capture many of the words in the corpus that are not part of CALD, including names of people (*Darwin*), technical words (*axon*), but also some words such as *later* and *easiest* that one would expect to be present in a thesaurus. Note that the reason words such as *later* and *easiest* are excluded in CALD is *not* because CALD lists only root word forms (related word forms are encoded in the database). These words appear to be genuine omissions in CALD that the concept–topic model handled by learned topics.

We also ran a simulation with the concept–topic model using the intersection vocabulary that includes words only present in both the TASA corpus and CALD database. Fig. 8B shows some example topics learned by the model. By definition, all words shown in these topics are members of some concepts so the concept–topic model is able to explain these words by first selecting a concept and then a word from a concept. These learned topics focus on word correlations that are not currently captured by the concepts. For example, the

(A)

word	prob.
darwin	0.049
charles	0.024
evolution	0.012
son	0.006
galapagos	0.006
lamarck	0.006
beagle	0.004
england	0.003
flytrap	0.003
malthus	0.003
wallace	0.003
alfred	0.003
geological	0.003
jacques	0.003
later	0.003

word	prob.
newton	0.084
galileo	0.027
isaac	0.018
later	0.013
newtons	0.010
inertial	0.008
permanganate	0.007
fleming	0.006
straight-line	0.006
huygens	0.004
italian	0.004
newtonian	0.004
alexander	0.004
tabletop	0.004
rectilinear	0.003

word	prob.
axon	0.012
fulcrum	0.011
dendrites	0.010
permanganate	0.008
acetylcholine	0.007
axons	0.007
nadph	0.005
parasympathetic	0.005
riverwood	0.005
easiest	0.004
inhibitory	0.004
cutter	0.004
effectors	0.004
energized	0.004
parkman	0.004

word	prob.
paleozoic	0.071
mesozoic	0.069
carbonyl	0.051
precambrian	0.043
cenozoic	0.041
cambrian	0.030
thermonuclear	0.024
aldehyde	0.019
quaternary	0.019
aldol	0.019
ketones	0.016
alvin	0.015
quats	0.013
coatings	0.012
flood-hazard	0.012

(B)

word	prob.
universe	0.094
galaxy	0.075
galaxies	0.062
milky	0.039
nebula	0.015
cosmic	0.013
billions	0.011
spiral	0.007
interstellar	0.007
resembles	0.006
acquiring	0.005
nebulae	0.005
galactic	0.004
static	0.004
accustomed	0.003

word	prob.
hypothesis	0.145
scientific	0.063
scientist	0.043
hypotheses	0.041
educated	0.011
suggested	0.010
outcome	0.009
suggests	0.008
verified	0.007
wrong	0.007
suggestions	0.005
duplicate	0.005
incorrect	0.005
searches	0.005
suggest	0.004

word	prob.
acres	0.023
farmer	0.019
swirling	0.008
cornfield	0.007
plowing	0.007
ranchers	0.006
differed	0.005
well-preserved	0.005
energetically	0.004
feasible	0.004
interlocking	0.004
splinter	0.004
tumbling	0.004
bewildering	0.003
buttocks	0.003

word	prob.
pollution	0.148
pollutants	0.053
era	0.034
chemicals	0.029
large	0.020
factories	0.018
smog	0.018
polluted	0.013
automobiles	0.013
fulcrum	0.011
sulfur	0.011
amounts	0.011
unfit	0.010
dumping	0.009
areas	0.009

Fig. 8. Examples of learned topics for the CTM model. Panel (A) illustrates a simulation using the union vocabulary that includes words that are part of the TASA corpus but are not part of the CALD vocabulary (these words are shown in bold). Panel (B) illustrated topics from a simulation on the intersection vocabulary that only includes words present in both TASA and CALD.

first two words in the left-most topic, *universe* and *galaxy*, are clearly related but are not members of the same concept. Similarly, *hypothesis* and *scientific* as well as *pollution* and *chemicals* are word pairs that are not members of the same concept but often co-occur in documents. The learned topics can be used to capture such correlations.

7. Expanding existing concepts with new words

One of the biggest disadvantages of utilizing human-defined knowledge databases is the large amount of manual effort involved to build and update the knowledge database. For

example, the lexicographers of the CALD database have to continually update the concepts to include words that have changed meaning or to insert entirely new concepts. In addition, the CALD database has to be checked for human errors, which might be difficult to detect manually. One way to test the utility of the concept model is to see whether it can automatically identify omissions within human-defined concepts, that is, words that should be in a concept but have been omitted. In the previous section, we showed how such words can become part of learned topics. In this section, we tested whether a concept–topic model could learn to expand existing concepts with new words that appear to have been omitted from concepts.

In our simulation approach, we removed selected words from the CALD concepts and tested how well a concept–topic model could use the TASA corpus to identify which concept they should be associated with. We only evaluate the concept–topic model with no learned topics (i.e., $T = 0$) on this concept recovery task (we expect that the hierarchical concept–topic model gives similar results). As a baseline method, we could compare the model against a number of existing models such as LDA or LSA. For simplicity, we focus here on LSA. We computed the singular value decomposition of the document word co-occurrence matrix for the TASA corpus and projected the terms onto N-dimensional concept space. Given a test word, we return a ranking of the concepts determined by the average distance from the test word to the m closest words in the concept. Distance is measured using cosine similarity. We experimented with different values of N and m, and report our results for $N = 150$ and $m = 5$ (results were relatively insensitive to the exact values used).

We compiled a list of 152 terms from the corpus where each term was a member of only one concept. The concept had to be well represented by the corpus, that is, at least 60% of the words in the concept were present in the corpus. Furthermore, the term had to be a significant member of the concept, that is, the term had the highest frequency in the corpus among all the terms in the same human-defined concept. In the concept–topic model, if a word is not included in the set of concepts to begin with, the model will be unaware of it. So for the purposes of this experiment, we ''removed'' a word by placing it in all 2,183 concept sets—in effect this tells the model that the word exists but gives the model no clue about which concept it belongs to. After training the model, we can simply count how often a word is assigned to a particular concept (via the z assignments) to produce a ranked list of concepts given a word.

Fig. 9 shows an example of the rankings returned by the concept–topic model for three test words. For each removed word, the figure shows the top five ranked concepts. We label each concept with one of four letters: M(atch) indicates a match to the target concept, P(arent) indicates a concept on the path from the root concept to the target concept, C(hild) indicates a concept in the subtree rooted at the target concept, and O(ther). In the example, the model is able to rank the target concept as the first or second ranked concept but even the highly ranked mismatching concepts are often quite reasonable target concepts. For example, the word *soot* has strong associations with the concept CLEANING AND TIDYING PLACES AND THINGS, as well as DIRT UNTIDINESS, which are semantically related to the word *soot*, but which did not originally contain the removed word.

Removed Word	Ranked Concepts
soot	
	(O) CLEANING AND TIDYING PLACES AND THINGS
	(M) PRODUCTS OF COMBUSTION
	(O) DIRT AND UNTIDINESS
	(O) BUILDINGS: NAMES AND TYPES OF
	(O) ENVIRONMENTAL ISSUES
insects	
	(C) INSECTS
	(M) INSECT NAMES
	(O) SOCIETY
	(O) IMPROVING FERTILITY AND PEST CONTROL
	(P) PLANTS AND ANIMALS
directions	
	(M) PLACES AND LOCATIONS
	(O) EMITTING AND CASTING LIGHT
	(O) POINTS OF THE COMPASS
	(O) SPORTS, GAMES AND PASTIMES
	(O) PAYING ATTENTION AND BEING CAREFUL

(M = match; C = child; P = parent; O = other)

Fig. 9. Example of rankings by concept–topic model in word recovery task.

Note that the concept–topic model is also able to identify multiple meanings for a given word. For example, the word *directions* can refer to north, south, east and west (KOINTS OF THE KOMPASS) as well as a set of instructions (PLACES AND LOCATIONS). Also note that many words in the CALD concepts are classified according to their definition (e.g., soot is a product of combustion) rather than their descriptive qualities (e.g., soot causes dirtiness and soot is an environmental issue).

Table 1 shows the overall results for the concept–topic model and latent semantic analysis. The table shows the probability that a concept is ranked in the top K returned concepts (precision), as a target concept, parent concept, or child concept. The results show that the concept–topic model outperforms the latent semantic analysis approach in this concept-recovery task. The concept–topic model is often able to rank the target concept (out of 2,183 concepts) in the top 10 or 20. For both the concept–topic models and the latent semantic analysis approach, the parent or child of the target concept also often appear (more than expected by chance) in the top 10 or 20 concepts, indicating that these models are able to recover more specific as well as more general concepts related to the novel word.

8. Discussion

Although most of the earlier work on topic modeling is purely data driven, in that no human knowledge is used in learning the topic model, there are some exceptions. Boyd-Graber, Blei, and Zhu (2007) develop a topic modeling framework that combines

Table 1
Precision results for the concept-recovery task

	LSA	Concept–Topic
Top 10 concepts		
Target	.45	.55
Parent	.22	.15
Child	.05	.04
Top 20 concepts		
Target	.53	.57
Parent	.30	.26
Child	.05	.05

human-derived linguistic knowledge using unsupervised topic models for the purpose of word–sense disambiguation. Andrzejewski, Zhu, and Craven (2009) recently introduced an iterative topic modeling process where a human can inspect the topics and specify which words should have high probability in a topic and which words should not appear together in a topic. By replacing the multinomial distribution over words in a topic with a Dirichlet forest prior, the knowledge expressed by a human can be taken into account in the next iteration of the topic modeling process. Wei and Croft (2007) use manually built topics using documents and categories from the Open Directory Project for information retrieval. The manual topics are built by aggregating documents for selected categories and obtaining probability distributions by normalizing the word counts of the associated documents.

Topic modeling has also been used for finding mappings between ontology pairs (Spiliopoulos, Vouros, & Karkaletsis, 2007). The work of Ifrim and Weikum (2006) and Bundschus, Dejori, Yu, Tresp, and Kriegel (2008) combines topics and concepts for the purposes of text classification. Our framework is somewhat more general in that we not only improve the quality of making predictions on text data by using prior human concepts but also are able to make inferences in the reverse direction about concept words and concept hierarchies given data. In addition, our concept–topic models do not require labeled data. Although topic modeling has also been used to semi-automatically build taxonomies from data (e.g., Dietz & Stewart, 2006; Zavitsanos, Paliouras, Vouros, & Petridis, 2007), these approaches do not make use of existing ontologies.

There is also a significant amount of prior work on using data to help with ontology construction, evaluation, and document tagging, such as learning ontologies from text data (e.g., Maedche & Staab, 2001), methodologies for evaluating how well ontologies are matched to specific text corpora (Alani & Brewster, 2006; Brewster, Alani, Dasmahapatra, & Wilks, 2004), and systems for tagging documents with semantic concepts using word-level matching techniques (Dill et al., 2003). Our work is broader in scope in that we propose general-purpose probabilistic models that combine concepts and topics within a single framework, allowing us to use the data to make inferences about how documents and concepts are related (for example). It should be noted that in the work reviewed in this paper, we do not explicitly investigate techniques for modifying an ontology in a data-driven manner (e.g., adding/deleting words from concepts

or relationships among concepts)—however, the framework we propose could certainly be used as a basis for exploring such ideas.

There are several potentially useful directions in which the hierarchical concept–topic model can be extended. One interesting extension to try is to substitute the Dirichlet prior on the concepts with a Dirichlet process prior, where each concept will now have a potentially infinite number of children, a finite number of which are observed at any given instance (e.g., Teh et al., 2006). When we do a random walk through the concept hierarchy to generate a word, we now have an additional option to create a child topic and generate a word from that topic. There would be no need for the switching mechanism as data-driven topics are now part of the concept hierarchy. Such a model would allow us to add new topics to an existing concept set hierarchy and could potentially be useful in building a recommender system for updating concept ontologies.

An alternative direction to pursue would be to introduce additional machinery in the generative model to handle different aspects of transitions through the concept hierarchy. In HCTM, we currently learn one set of path correlations for the entire corpus (captured by the Dirichlet parameters τ in HCTM). It would be interesting to introduce another latent variable to model multiple path correlations. Under this extension, documents from different genres can learn different path correlations (similar to the work of Boyd-Graber et al., 2007). For example, scientific documents could prefer to utilize paths involving scientific concepts, and humanities concepts could prefer to utilize a different set of path correlations when they are modeled together. A model of this type would also be able to make use of class labels of documents if available.

9. Conclusions

We have proposed a probabilistic framework for combining data-driven topics and semantically rich human-defined concepts. We first introduced the concept–topic model, a straightforward extension of the topic model, to utilize human-defined semantic concepts in the topic modeling framework. The model represents documents as a mixture of topics and concepts, thereby allowing us to describe documents using the semantically rich concepts. We further extended this model with the hierarchical concept–topic model where we incorporate the concept hierarchy into the generative model by modeling the parent–child relationship in the concept hierarchy.

Our experimental results show that the semantic concepts significantly improve the quality of the resulting models. Modeling concepts and their associated hierarchies appears to be particularly useful when there are limited training data—the hierarchical concept–topic model has the best predictive performance overall in this regime. We view the current set of models as a starting point for exploring more expressive generative models that can potentially have wide-ranging applications, particularly in areas of document modeling and tagging, ontology modeling and refining, and information retrieval.

In addition, these models are useful to expand the current cognitive science framework to characterize human learning of semantic information. Many existing models in cognitive

science explain how a human learner extracts semantic information from only a single source of information: statistical co-occurrence information between words and documents. The current set of models suggests that the learning process in such models can be enhanced when additional background information is available. For example, a human learner might already be familiar with certain concepts (and the relations between concepts) and the exposure to (new) statistical information such as word–document co-occurrences serves to refine existing concepts or perhaps learn new ones.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation under Award Number IIS-0083489, and by the Office of Naval Research under Award Number N00014-08-1-1015. We are extremely grateful to three anonymous reviewers, and to Danielle McNamara and Simon Dennis, for their very helpful comments on an earlier version of this article. We thank America Holloway for assistance in providing the results in Section 7.

References

Alani, H., & Brewster, C. (2006). Metrics for ranking ontologies. *4th International EON Workshop, 15th International World Wide Web Conference*. New York: ACM.

Andrzejewski, D., Zhu, X., & Craven, M. (2009). Incorporating domain knowledge into topic modeling via Dirichlet forest priors. *The 26th International Conference on Machine Learning (ICML)*. New York: ACM.

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2003). Hierarchical topic models and the nested Chinese restaurant process. In S. Thrun, L. Saul, & B. Scholkopf (Eds.), *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press.

Blei, D., & Jordan, M. (2003). Modeling annotated data. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 127–134). New York: ACM.

Blei, D., & Lafferty, J. (2006). Correlated topic models. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (pp. 147–154). 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, D., Blei, D., & Zhu, X. (2007). A topic model for word sense disambiguation. *Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1024–1033). New York: ACM.

Brewster, C., Alani, H., Dasmahapatra, S., & Wilks, Y. (2004). Data driven ontology evaluation. *International Conference on Language Resources and Evaluation*. Paris, France.

Brown, P. F., deSouza, P. V., Mercer, R. L., Della Pietra, V. J., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467–479.

Bundschus, M., Dejori, M., Yu, S., Tresp, V., & Kriegel, H. (2008). Statistical modeling of medical indexing processes for biomedical knowledge information discovery from text. *Proceedings of the 8th International Workshop on Data Mining in Bioinformatics (BIOKDD)*. New York: ACM.

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.

Chater, N., & Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344.

Chemudugunta, C., Holloway, A., Smyth, P., & Steyvers, M. (2008a). Modeling documents by combining semantic concepts with unsupervised statistical learning. *7th International Semantic Web Conference* (pp. 229–244). Berlin: Springer-Verlag.

Chemudugunta, C., Smyth, P., & Steyvers, M. (2008b). Combining concept hierarchies and statistical topic models. *ACM 17th Conference on Information and Knowledge Management*. New York: ACM.

Chemudugunta, C., Smyth, P., & Steyvers, M. (2008c). *Text modeling using unsupervised topic models and concept hierarchies*. Technical Report. Available at: http://arxiv.org/abs/0808.0973. Accessed on May 16, 2010.

Cutting, D. R., Karger, D., Pedersen, J. O., & Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 318–329). New York: ACM Press.

Dennis, S. (2004). An unsupervised method for the extraction of propositional information from text. *Proceedings of the National Academy of Sciences*, *101*, 5206–5213.

Dietz, L., & Stewart, A. (2006). Utilize probabilistic topic models to enrich knowledge bases. *Proceedings of the E SWC 2006 workshop on mastering the gap: From information extraction to semantic representation*.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J., & Zien, J. (2003). SemTag and seeker: Bootstrapping the semantic web via automated semantic annotation. *Proceedings of the 12th international conference on World Wide Web* (pp. 178–186). New York: ACM.

Fellbaum, C. (Ed.) (1998). *WordNet, an electronic lexical database*. Cambridge, MA: MIT Press.

Fiser, J., & Aslin, R. N. (2005). Encoding multi-element scenes: Statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General*, *134*, 521–537.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments*, *8*, 111–129.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*, 2nd ed. London: Chapman & Hall.

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*. In L. K. Saul (Ed.), *Advances in neural information processing 17* (pp. 537–544). 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., Speer, R., & Alonso, J. (2007). ConceptNet 3: A flexible, multilingual semantic network for common sense knowledge. *Proceedings of Recent Advances in Natural Language Processing*.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval* (pp. 50–57). New York: ACM Press.

Ifrim, G., & Weikum, G. (2006). Transductive learning for text classification. *10th European conference on principles and practice of knowledge discovery in databases* (pp. 223–234). Berlin, Germany.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1–37.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., & Saarela, A. (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.

Lagus, K., Honkela, T., Kaski, S., & Kohonen, T. (1999). WEBSOM for textual data mining. *Artificial Intelligence Review*, *13*, 5–6. 345–364.

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.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Lenat, D. B., & Guha, R. V. (1989). *Building large knowledge-based systems: Representation and inference in the Cyc project*. Reading, MA: Addison-Wesley.

Li, W., Blei, D., & McCallum, A. (2007). Nonparametric Bayes pachinko allocation. *Conference on Uncertainty in Artificial Intelligence (UAI)*. Corvallis, OR: AUAI Press.

Maedche, A., & Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems*, *16*(2), 72–79.

McCallum, A., Nigam, K., & Ungar, L. H. (2000). Efficient clustering of high-dimensional data sets with application to reference matching. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 169–178). New York: ACM Press.

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.

Mei, Q., Shen, X., & Zhai, C. (2007). Automatic labeling of multinomial topic models. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 490–499). New York: ACM Press.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to Word Net: An on line lexical database. *International Journal of Lexicography*, *3*(4), 235–244.

Mimno, D. M., Li, W., & McCallum, A. (2007). Mixtures of hierarchical topics with pachinko allocation. *International Conference on Machine Learning (ICML)* (pp. 633–640). Corvallis, OR.

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

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*, 387–391.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. Available at: http://www.usf.edu/FreeAssociation/.

Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. *Springer Lecture Notes in Computer Science (LNCS) series—IEEE international conference on intelligence and security informatics*. Berlin: Springer-Verlag.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*, 127–162.

Newport, E. L., Hauser, M. D., Spaepen, G., & Aslin, R. N. (2004). Learning at a distance: II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology*, *49*, 85–117.

Panton, K., Mtuszek, C., Lenat, D., Schneider, D., Witbrock, M., Siegel, N., & Shepard, B. (2006). Common sense reasoning—From Cyc to intelligent assistant. *Lecture Notes in Computer Science*, *3864*, 1–31.

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.

Roget, P. M. (1911). *Roget's thesaurus of English words and phrases*. New York: Thomas Y. Crowell.

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.

Spiliopoulos, V., Vouros, G., & Karkaletsis, V. (2007). Mapping ontologies elements using features in a latent space. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 457–460). Washington DC: IEEE Computer Society.

Steyvers, M., & Griffiths, T. L. (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 427–448). Mahwah, NJ: Erlbaum.

Steyvers, M., & Griffiths, T. L. (2008). Rational analysis as a link between human memory and information retrieval. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects from rational models of cognition* (pp. 327–347). Oxford, England: Oxford University Press.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In W. Kim, R. Kohavi, J. Gehrke, & W. DuMouchel (Eds.), *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 306–315). New York: ACM.

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

Wallach, H. (2006). Topic modeling: Beyond bag-of-words. In W. Cohen & A. Moore (Eds.), *Proceedings of the 23rd International Conference on Machine Learning* (pp. 977–984). Pittsburgh, PA.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In A. P. Danyluk, L. Bottou, & M. L. Littman (Eds.), *Proceeding of the 26th International Conference on Machine Learning* (pp. (1105–1112.) New York: ACM.

Wang, C., Blei, D., & Heckerman, D. (2008). Continuous time dynamic topic models. In D. McAllester and A. Nicholson (Eds.), *Uncertainty in artificial intelligence* (pp. 579–586). Corvallis, OR: AUAI Press.

Wei, X., & Croft, W. B. (2007). Investigating retrieval performance with manually-built topic models. *Proceedings of the 8th Large-Scale Semantic Access to Content (Text, Image, Video and Sound) Conference (RIAO'07)*. Paris, France.

Yu, C., Ballard, D. H., & Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science*, *29*, 961–1005.

Zavitsanos, E., Paliouras, G., Vouros, G.A., & Petridis, S. (2007). Discovering subsumption hierarchies of ontology concepts for text corpora. *In proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence* (pp.402–408.) Washington, D.C.: IEEE Computer Society.