Modeling Documents by Combining Semantic Concepts with Unsupervised Statistical Learning

Chaitanya Chemudugunta[1], America Holloway[1], Padhraic Smyth[1], and Mark Steyvers[2]

[1] Department of Computer Science
University of California, Irvine, Irvine, CA
{chandra,ahollowa,smyth}@ics.uci.edu
[2] Department of Cognitive Science
University of California, Irvine, Irvine, CA
msteyver@uci.edu

Abstract. Human-defined concepts are fundamental building-blocks in constructing knowledge bases such as ontologies. Statistical learning techniques provide an alternative automated approach to concept definition, driven by data rather than prior knowledge. In this paper we propose a probabilistic modeling framework that combines both human-defined concepts and data-driven topics in a principled manner. The methodology we propose is based on applications of statistical topic models (also known as latent Dirichlet allocation models). We demonstrate the utility of this general framework in two ways. We first illustrate how the methodology can be used to automatically tag Web pages with concepts from a known set of concepts without any need for labeled documents. We then perform a series of experiments that quantify how combining human-defined semantic knowledge with data-driven techniques leads to better language models than can be obtained with either alone.

Key words: ontologies, tagging, unsupervised learning, topic models

1 Introduction

An important step towards a semantic Web is automated and robust annotation of Web pages and online documents. In this paper we consider a specific version of this problem, namely, mapping of an entire document or Web page to concepts in a given ontology. To address this problem we propose a probabilistic framework for combining ontological concepts with unsupervised statistical text modeling. Here, and throughout, we use the term ontology to refer to *simple ontologies* [1] which are collections of human-defined concepts usually with a hierarchical structure. In this paper we focus on the simplest aspect of these ontologies, namely the ontological concepts and associated vocabulary (and to a lesser extent the hierarchical relations between concepts). We focus our investigation on the overall feasibility of the proposed approach—given the promise of the results obtained in this paper, the next step will be to develop models that can leverage the richer aspects of ontological knowledge representation.

We use statistical topic models (also known as latent Dirichlet allocation models [2, 3]) as the underlying quantitative modeling framework. *Topics* from statistical models and *concepts* from ontologies both represent "focused" sets of words that relate to some abstract notion—this similarity is the key idea we exploit in this paper. As an example, Table 1 lists some of the 204 words that have been manually defined as part of the concept FAMILY in the Cambridge International Dictionary of English (CIDE: more details on this ontology are provided later in the paper). The second column is a topic, also about families, that was learned automatically from a text corpus using a statistical topic model.

FAMILY Concept	FAMILY Topic	
beget	family	(0.208)
birthright	child	(0.171)
brood	parent	(0.073)
brother	young	(0.040)
children	boy	(0.028)
distantly	mother	(0.027)
dynastic	father	(0.021)
elder	school	(0.020)

Table 1. CIDE FAMILY concept and learned FAMILY topic

The numbers in parentheses are the probabilities that a word will be generated conditioned on the learned topic—these probabilities sum to 1 over the entire vocabulary of words, specifying a multinomial distribution. The concept FAMILY in effect puts probability mass 1 on the set of 204 words within the concept, and probability 0 on all other words. The topic multinomial on the other hand could be viewed as a "soft" version of this idea, with non-zero probabilities for all words in the vocabulary—but significantly skewed, with most of the probability mass focused on a relatively small set of words.

Many of the existing methods for semantic annotation of Web pages are focused on specific entity-tagging tasks, using a variety of natural language processing (NLP), information extraction (IE), and statistical language modeling techniques (e.g., [4–6]). A well-known semantic annotation system of this type is SemTag [7] which was built to annotate entity-rich web pages on a large scale. The main differences between this past work and our approach are that we map all words in a document, not just entities, onto a set of ontological concepts, we learn a probabilistic model over words and concepts, and we use an entirely unsupervised approach without any need for supervised labeling.

There has also been prior work that combines ontological concepts and data-driven learning within a single framework, such as using concepts as pre-processing for text modeling [8, 9], using word-concept distributions as a form of background knowledge to improve text-classification [10], and combining human-derived linguistic knowledge with topic-based learning for word-sense disambiguation [11]. There has also been work on developing quantitative methods for evaluating how well ontologies fit specific text corpora [12, 13] as well as a significant amount of research on ontology learning from data. Our work is different from all of this prior work in that we propose probabilistic models that combine concepts and data-driven topics within a single general framework, allowing (for example) the data to enable inferences about the concepts.

We begin the paper by reviewing the general ideas underlying statistical topic modeling and then show how these techniques can be directly adapted for the purposes of combining semantic concepts with text corpora. In the remainder of the paper we illustrate how the resulting models can be used to automatically tag words in Web pages and map each word into an ontological concept taking into account the context of the document. Additionally, we describe a set of quantitative experiments that evaluate the quality of the models when viewed as language models. We conclude that combining semantic concepts and data-driven topic learning opens up new opportunities and applications that would not be possible using either technique alone.

2 A Review of Statistical Topic Models

The latent Dirichlet allocation (LDA) model, also referred to as the topic model, is a state-of-the-art unsupervised learning technique for extracting thematic information from large document sets [2, 3]. In this section we briefly review the fundamental ideas behind this model since it provides the basis for our approach later in the paper.

Let $\{w_1, \ldots, w_V\}$ be the set of unique words in a corpus, where V is the size of the vocabulary. Each document in the corpus is represented as a "bag of words", namely a sparse vector of length V where component i contains the number of times word i occurs in the document.

HEALTH CARE		FARMING	
health	(0.064)	farm	(0.081)
care	(0.058)	crop	(0.027)
plan	(0.047)	cow	(0.018)
cost	(0.043)	field	(0.015)
insurance	(0.042)	corn	(0.015)
benefit	(0.032)	food	(0.012)
converage	(0.023)	bean	(0.010)
pay	(0.020)	cattle	(0.010)
program	(0.013)	market	(0.010)

Table 2. Two example topics learned from a large corpus

A topic $z_j, 1 \leq j \leq T$ is represented as a multinomial probability distribution over the V words, $p(w_i|z_j), \sum_i^V p(w_i|z_j) = 1$. Simulating n words from a topic is analogous to throwing a die n times except that instead of 6 equiprobable outcomes on each throw we have V possible outcomes (where V can be on the order of 100,000 in practice) and the probabilities of individual outcomes (the words) may be significantly non-uniform. Table 2 shows two example topics that were learned from a large corpus (more details on learning below). The topic names are generally assigned manually. If we simulate data from one of these topics, the high probability words (shown in the figure) will occur with high frequency. A topic, in the form of a multinomial distribution over a vocabulary of words, can in a loose sense be viewed as a probabilistic representation of a semantic concept.

The topic model assumes that words in a document arise via a two-stage process: words are generated from topics and topics are generated by documents. More formally the distribution of words given a document, $p(w_i|d)$, is modeled as a mixture over topics:

$$p(w_i|d) = \sum_{j=1}^{T} p(w_i|z_j)p(z_j|d). \tag{1}$$

The topic variable z plays the role of a low-dimensional representation of the semantic content of a document.

Intuitively we can imagine simulating n words in a document by repeating the following steps n times: first, sample a topic z_j from the topic-document distribution $p(z|d)$, and then, given a topic z_j, sample a word from the corresponding word-topic distribution $p(w|z_j)$. For example, imagine that we have the following 5 topics with corresponding probability distributions over words: *earthquake*, *disaster response*, *international politics*, *China*, and *Olympic Games*. We could then represent individual documents as weighted combinations of this "basis set" of topics, e.g., one document could be a mixture of words from the topics *earthquake*, *disaster response*, and *China*, while another document could be a mixture of words from *China*, *international politics*, and *Olympic Games*.

By allowing documents to be composed of different combinations of topics, a topic model provides a more flexible representation of document content than clustering where each document is assumed to have been generated by a single cluster. Topics can also be considered a more natural representation for document content than the technique of latent semantic analysis (LSA) [14] since the multinomial basis of the topic model is better suited to predicting word counts than the inherently real-valued/least-squares framework that underlies LSA. A number of studies have shown that topic models provide systematically better results in document modeling and prediction compared to LSA ([15], [16]).

In the standard topic-modeling framework the word-topic distribution $p(w|z)$ and topic-document distributions $p(z|d)$ are learned in a completely unsupervised manner, without any prior knowledge of what words are associated with topics or what topics are associated with individual documents. The statistical estimation technique of Gibbs sampling is widely used [3]: starting with random assignments of words to topics, the algorithm repeatedly cycles through the words in the training corpus and samples a topic assignment for each word using the conditional distribution for that word given all other current word-topic assignments (see Appendix 1 for more details). After a number of such iterations through all words in the corpus (typically on the order of 100) the algorithm reaches a steady-state. The word-topic probability distributions can be estimated from the word-topic assignments. It is worth noting that topic model learning results in assignments of topics to each word in the corpus. This in turn directly enables "topic-tagging" of words, sentences, sections, documents, groups of documents, etc., a feature we will leverage later in this paper.

3 Semantic Concepts and Statistical Topic Modeling

We now return to the topic of concepts within ontologies and show how the statistical topic modeling techniques of the previous section can leverage text corpora to "overlay" probabilities on such concepts. As mentioned in the introduction, in this paper we focus on a simple aspect of ontological knowledge, namely sets of words associated with concepts.

FARMING & FORESTRY		EARTH & OUTER SPACE	
crops	(0.135)	earth	(0.226)
plant	(0.076)	sky	(0.107)
grow	(0.050)	space	(0.082)
land	(0.040)	sun	(0.066)
fertilizers	(0.038)	scientists	(0.046)
soil	(0.037)	planets	(0.033)
earth	(0.034)	universe	(0.033)
farming	(0.034)	stars	(0.032)

Table 3. Two example concepts from the CIDE thesaurus

Assume that we have been given a set of C human-defined concepts, where each concept c_j consists of a finite set of N_j unique words, $1 \leq j \leq C$. We also have available a corpus of documents such as Web pages. We propose to merge these two sources of information (concepts and documents) using a framework based on topic modeling. For example, we might be interested in "tagging" documents with concepts from the ontology, but with little or no supervised labeled data available (note that the approach we describe below can be easily adapted to include labeled documents if available). One way to approach this problem would be to assume a model in the form of a topic model, i.e.,

$$p(w_i|d) = \sum_{j=1}^{C} p(w_i|c_j)p(c_j|d). \tag{2}$$

which is the same as Equation 1 but where we have replaced topics z with concepts c. We will refer to this type of model as the *concept model* throughout the paper. In the concept model the words that belong to a concept are defined by a human a priori (e.g., as part of an ontology) and are limited (typically) to a small subset of the overall vocabulary. In contrast, in a topic model, all words in the vocabulary can be associated with any particular topic but with different probabilities.

In Equation 2 above, the unknown parameters of the concept model are the word-concept probabilities $p(w_i|c_j)$ and the concept-document probabilities $p(c_j|d)$. Our goal (as in the topic model) is to estimate these from an appropriate corpus. Note for example that the probabilities $p(c_j|d)$ would address the afore-mentioned tagging problem, since each such distribution tells us the mix of concepts c_j that a document d is represented by.

We can use a modified version of statistical topic model learning algorithm to infer both $p(w_i|c_j)$ and $p(c_j|d)$. The process is to simply treat concepts as "topics with constraints," where the constraints consist of setting words that are not a priori mentioned in a concept to have probability 0, i.e., $p(w_i|c_j) = 0, w_i \notin c_j$. We can use Gibbs sampling to assign concepts to words in documents, using the same sampling equations as used for assigning topics to words in the topic model, but with the additional constraint that a word can only be assigned to a concept that it is associated with in the ontology[3]. Other than the constraint restriction, the learning algorithm is exactly the same as in standard learning of topic models, and the end result is that each word in the corpus is assigned to a concept in the ontology. In turn, these assignments allow us to directly estimate the terms of interest in Equation 2 above. To estimate $p(w_i|c_j)$ for a particular concept c_j we count how many words in the corpus were assigned by the sampling algorithm to concept c_j and normalize these counts (and typically also smooth them) to arrive at the probability distribution $p(w_i|c_j)$. To estimate $p(c_j|d)$ for a particular document d, we count how many times each concept is assigned to a word in document d and again normalize and smooth the counts to obtain $p(c_j|d)$. Table 3 shows an example of a set of learned probabilities for words (ranked highest by probability) for two different concepts from the CIDE ontolgy, after training on the TASA corpus (more details on ontologies and data sets are provided later).

The important point to note here is that we have defined a straightforward way to "marry" the qualitative information in sets of words in human-defined concepts with quantitative data-driven topics. The learning algorithm itself is not innovative, but the application is innovative in that it combines two sources of information (concepts from ontologies and statistical learning) that to our knowledge have not been combined in any general framework in prior work. We can use the learned probabilistic representation of concepts to map new documents into concepts within an ontology, and we can use the semantic concepts to improve the quality of data-driven topic models. We will explore both of these ideas in more detail in later sections of the paper.

There are numerous variations of the concept model framework that can be explored—we investigate some of the more obvious extensions below. For example, a baseline model is one where the word-concept probabilities $p(w_i|c_j)$ are defined to be uniform for all words within a concept. A related model is one where the word-concept probabilities are available a priori as part of the concept definition, e.g., where documents are provided with each concept allowing for empirical word-concept probabilities to be estimated. For both of these models, Gibbs sampling is still used as before to infer the word-concept assignments and the concept-document probabilities, but the $p(w|c)$ probabilities are held fixed and not learned. We will refer to these two models as ConceptU (concept-uniform) and ConceptF (concept-fixed) and use ConceptL (concept-learned) to refer to the more general concept model described earlier where the $p(w|c)$ probabilities are learned from the corpus.

Human-generated concepts not only come with words associated with concepts but are also often arranged in a hierarchical structure such as a concept tree, where each node is a concept with a set of associated words. A simple way to incorporate this hierarchical information is to propagate the words upwards in the concept tree, so that an internal concept node is associated with its own words and all the words associated with its children. When we use this propagation technique for representing the word-

[3] An alternative approach, not explored in this paper, would be to use the concept words to build an informative prior on topics rather than using them as a hard constraint. Under such an approach, each concept could be associated with any word in the corpus leading to significant computational demands since large ontologies could have tens of thousands of concepts.

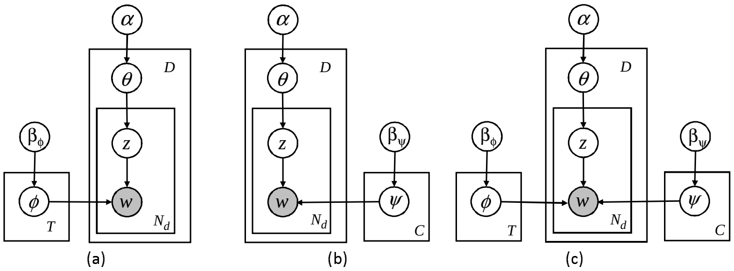


Fig. 1. Graphical models for (a) Topic model, (b) Concept model, and (c) Concept-topic model

concept associations, we will refer to this by adding an "H" to the name of the learned model, e.g., ConceptLH, ConceptFH, etc.

Finally, a natural further extension of the model is to allow for incorporation of unconstrained data-driven topics alongside the concepts. This can be achieved by simply allowing the Gibbs sampling procedure to either assign a word to a constrained concept or to one of the unconstrained topics (see Appendix 1). In such a model a document is represented by a mixture over C concepts and T topics, allowing the model to use additional data-driven topics to represent themes that are not well-represented in the set of concepts in the ontology. We will in general refer to such models as concept-topic models and specific variations by ConceptL+Topics, ConceptLH+Topics etc.

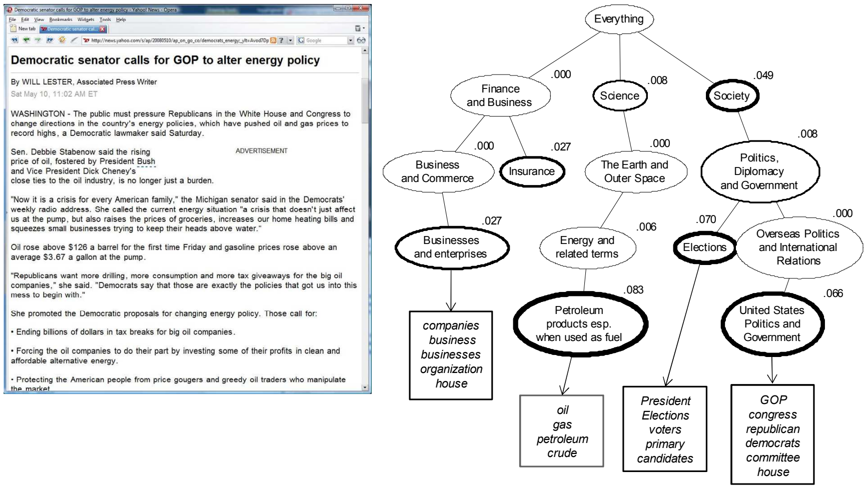
Figure 1 shows a graphical model representation of the various models, including the standard topic model, the concept model, and the concept-topic model. Here, ϕ, ψ and θ represent word-topic, word-concept and topic-document/concept-document multinomial distributions respectively. β_ϕ, β_ψ and α represent the Dirichlet priors on ϕ, ψ and θ respectively. Further details on sampling equations for all of the model variants are provided in Appendix 1.

4 Concept Sets and Text Data

The experiments in this paper are based on one large text corpus and two different knowledge bases. For the text corpus, we used the Touchstone Applied Science Associates (TASA) dataset [14]. This corpus consists 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 9 different educational topics. In this paper, we focus on the documents classified as SCIENCE and SOCIAL STUDIES, consisting of $D = 5356$ and $D = 10,501$ documents and 1.7M and 3.4M word tokens respectively.

The first set of concepts we used was the Open Directory Project (ODP), a human-edited hierarchical directory of the web (available at http://www.dmoz.org). The ODP database contains descriptions and urls on a large number of hierarchically organized topics. We extracted all the topics in the SCIENCE subtree, which consists of $C = 10,817$ nodes after preprocessing. The top concept in this hierarchy starts with SCIENCE and divides into concepts such as ASTRONOMY, MATH, PHYSICS, etc. Each of these topics divides again into more specific concepts with a maximum number of 11 levels. Each node in the hierarchy is associated with a set of urls related to the concept plus a set of human-edited descriptions of the site content. To create a bag of words representation for each node, we collected all the words in the textual descriptions and also crawled the urls associated with the node (a total of 78K sites). This led to a vector of word counts for each node.

The second source of concepts in our experiments was a thesaurus from the Cambridge International Dictionary of English (CIDE; www.cambridge.org/elt/cide). CIDE consists of $C = 1923$ hierarchically organized semantic categories. In contrast to other taxonomies such as WordNet [17], CIDE 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, etc). The hierarchy has up to 7 levels. The concepts vary in the number of the words with a median of 54 words and a maximum of 3074. Each word can be a member of multiple concepts, especially if the word has multiple senses.


Democratic senator calls for GOP to alter energy policy - Yahoo! News - Opera

File Edit View Bookmarks Widgets Tools Help

New tab | Democratic senator cal...

http://news.yahoo.com/s/ap/20080510/ap_on_go_co/democrats_energy;_ylt=Avod7Dp

Democratic senator calls for GOP to alter energy policy

By WILL LESTER, Associated Press Writer

Sat May 10, 11:02 AM ET

WASHINGTON - The public must pressure Republicans in the White House and Congress to change directions in the country's energy policies, which have pushed oil and gas prices to record highs, a Democratic lawmaker said Saturday.

Sen. Debbie Stabenow said the rising price of oil, fostered by President Bush and Vice President Dick Cheney's close ties to the oil industry, is no longer just a burden.

ADVERTISEMENT

"Now it is a crisis for every American family," the Michigan senator said in the Democrats' weekly radio address. She called the current energy situation "a crisis that doesn't just affect us at the pump, but also raises the prices of groceries, increases our home heating bills and squeezes small businesses trying to keep their heads above water."

Oil rose above $126 a barrel for the first time Friday and gasoline prices rose above an average $3.67 a gallon at the pump.

"Republicans want more drilling, more consumption and more tax giveaways for the big oil companies," she said. "Democrats say that those are exactly the policies that got us into this mess to begin with."

She promoted the Democratic proposals for changing energy policy. Those call for:

- Ending billions of dollars in tax breaks for big oil companies.

- Forcing the oil companies to do their part by investing some of their profits in clean and affordable alternative energy.

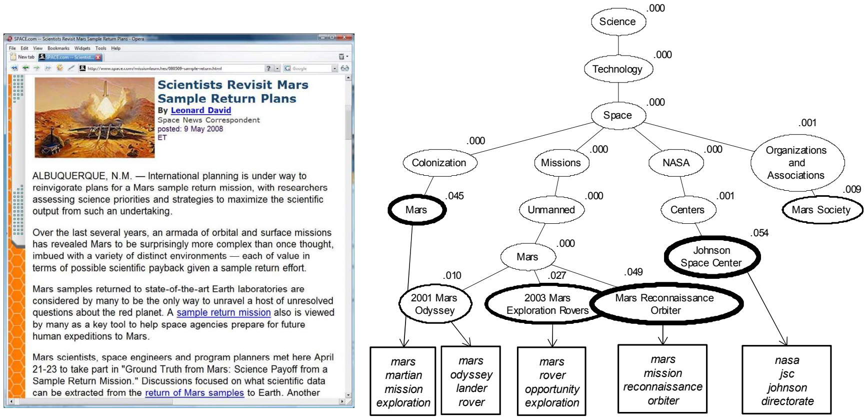
- Protecting the American people from price gougers and greedy oil traders who manipulate the market

Everything
.000
Finance and Business
.008
Science
.049
Society
.000
Business and Commerce
.027
Insurance
.000
The Earth and Outer Space
.008
Politics, Diplomacy and Government
.027
Businesses and enterprises
.006
Energy and related terms
.070
Elections
.000
Overseas Politics and International Relations
.083
Petroleum products esp. when used as fuel
.066
United States Politics and Government
companies business businesses organization house
oil gas petroleum crude
President Elections voters primary candidates
GOP congress republican democrats committee house

Fig. 2. Example of using the ConceptU model to automatically tag a Web page with CIDE concepts.

5 Tagging Documents with Concepts

One application of concept models is to tag documents such as Web pages with concepts from the ontology. 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 Web page, whole Web pages or even collections of


SPACE.com -- Scientists Revisit Mars Sample Return Plans - Opera
File Edit View Bookmarks Widgets Tools Help
New tab
SPACE.com -- Scientist...
http://www.space.com/missionlaunches/080509-sample-return.html
Google

Scientists Revisit Mars Sample Return Plans

By Leonard David
Space News Correspondent
posted: 9 May 2008
ET

ALBUQUERQUE, N.M. — International planning is under way to reinvigorate plans for a Mars sample return mission, with researchers assessing science priorities and strategies to maximize the scientific output from such an undertaking.

Over the last several years, an armada of orbital and surface missions has revealed Mars to be surprisingly more complex than once thought, imbued with a variety of distinct environments — each of value in terms of possible scientific payback given a sample return effort.

Mars samples returned to state-of-the-art Earth laboratories are considered by many to be the only way to unravel a host of unresolved questions about the red planet. A sample return mission also is viewed by many as a key tool to help space agencies prepare for future human expeditions to Mars.

Mars scientists, space engineers and program planners met here April 21-23 to take part in "Ground Truth from Mars: Science Payoff from a Sample Return Mission." Discussions focused on what scientific data can be extracted from the return of Mars samples to Earth. Another

Science
.000
Technology
.000
Space
.000
Colonization
.000
Missions
.000
NASA
.000
Organizations and Associations
.001
Mars
.045
Unmanned
.000
Centers
.001
Mars Society
.009
Mars
.000
Johnson Space Center
.054
2001 Mars Odyssey
.010
2003 Mars Exploration Rovers
.027
Mars Reconnaissance Orbiter
.049
mars martian mission exploration
mars odyssey lander rover
mars rover opportunity exploration
mars mission reconnaissance orbiter
nasa jsc johnson directorate

Fig. 3. Example of using the ConceptU model to automatically tag a Web page with ODP concepts.

Web pages. Figure 2 illustrates the effect of tagging a Web page with CIDE concepts using the ConceptU model. For the purpose of illustration, the six highest probability concepts along will their parents and ancestors are shown. The thickness of the ellipse encapsulating a concept node is proportional to the probability of the concept in the Web page. The rectangular boxes contain words from the Web page that were assigned to the corresponding concept in decreasing order of frequency. Figure 3 shows an example of tagging another Web page using the ConceptU model with concepts from the ODP ontology, with "Johnson Space Center" and "Mars Reconnaissance Orbiter" among the high probability concepts. For these tagging illustrations, we ran 1500 Gibbs sampling chains and each chain was run for 50 iterations after which a single sample was taken.

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. Example of tagging at the word level using the ConceptL model.

Figure 4 illustrates concept assignments to individual words in a TASA document with CIDE concepts. 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. 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 model, the distributions over concepts within a document are highly skewed such that most probability goes to only a small number of concepts. In the example document, the four most likely concepts cover about 50% of all words in the document.

The figure illustrates that the model correctly disambiguates words that have several conceptual interpretations. For example, the word *charged* has many different meanings and appears in 20 CIDE concepts. In the example document, this word is assigned to the physics concept which is a reasonable interpretation in this document context. Similarly, the ambiguous words *current* and *flow* are correctly assigned to the electricity concept.

6 Language Modeling Experiments

To quantitatively measure the quality of the concept models described in the earlier parts of the paper, we perform a set of systematic experiments that compare the quality of concept models and baselines. To do this we use standard techniques from language modeling that measure the predictive power of a model in terms of its ability to predict words in unseen documents.

6.1 Perplexity

Perplexity is widely used as a quantitative measure for comparing language models, e.g. [18]. It can be interpreted as being proportional to the distance (formally, the cross-entropy) between the word distribution learned by the model and the distribution of words in an unseen test document. Thus, lower scores are better since they indicate that the model's distribution is closer to that of the actual text. The perplexity of a test data set is defined as:

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

where $\mathbf{w}_{test}$ is the words in test documents, $\mathbf{w}_d$ are words in document d of the test set, $\mathcal{D}^{\text{train}}$ is the training set, and N_d is the number of words in document d

In the experiments that follow we partition the text corpus into disjoint training and test sets, with 90% of the documents being used for training and the remaining 10% for computing test perplexity. For each test document d, a randomly selected subset of 50% of the words in the document are assumed to be observed and used to estimate the document-specific parameters $p(c|d)$ and/or $p(z|d)$ via Gibbs sampling. Perplexity is then computed on the remaining 50% of the words in the document (a form of perplexity known as predictive-perplexity).

In our experiments below we use perplexity to evaluate the relative quality of different concept and concept-topic models. Although no single quantitative measure will necessarily provide an ideal measure of how well human concepts and a corpus are matched, we argue that perplexity scores have the appropriate behavior. In particular,

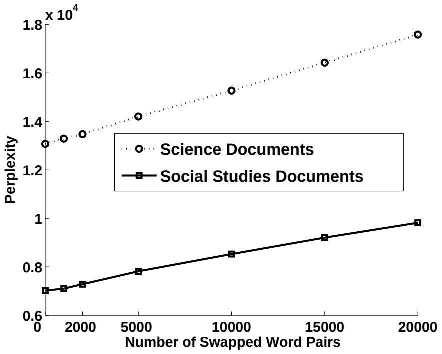


Fig. 5. Perplexity as a function of precision.

perplexity will be sensitive to both the precision and recall of a knowledge-base in relation to a corpus. Precision in this context should measure the semantic coherence of words within a concept and recall should be sensitive to how well the concepts cover a body of knowledge (e.g., as represented by a corpus) [12]. Therefore, as precision or recall increase we expect perplexity to decrease. We illustrate this (for precision) with a simulated experiment where we swap words randomly between CIDE concepts (to intentionally "corrupt" the concepts) and then measure the quality of the resulting concept model on the TASA corpus using the ConceptU model. As the number of words swapped increases (x-axis in Figure 5) the precision decreases, and the resulting perplexity very clearly reflects the deterioration in the quality of the concepts. Thus, perplexity appears to be a reasonable surrogate measure for more ontology-specific notions of quality such as precision.

6.2 General Perplexity Results across Models

We created a single $W = 33,635$ word vocabulary based on the 3-way intersection between the vocabularies of TASA, CIDE, and ODP. This vocabulary covers 89.9% of all of the word tokens in the TASA corpus and is the vocabulary that is used in all of the experiments reported in this paper. We also generated the same set of experimental results below using the union of words in TASA and CIDE and TASA and ODP, and found the same general behavior as with the intersection vocabulary. We report the intersection results below and omit the union results as they are essentially identical to the intersection results. A useful feature of using the intersection is that it allows us to evaluate two different sets of concepts (TASA and CIDE) on a common data set (TASA) and vocabulary, e.g., to evaluate which set of human-defined concepts better predicts a given set of text data. Note that selecting a predefined vocabulary (whether the intersection or the union) bypasses the important practical problem of modeling "out of vocabulary" words that may be seen in new documents. Although this is an important aspect of language modeling in general, in this paper our primary focus is on combining human defined concepts and data-derived topics.

Table 4 shows predictive perplexity scores for a variety of models using the TASA corpus with the CIDE or ODP concepts. In terms of general trends, there is a systematic reduction in perplexity scores as more corpus-specific information is combined with the concepts. The concept models with uniform distributions (ConceptU) have relatively high perplexity scores, indicating that a uniform distribution over concept terms are a poor fit to the data as one would expect. Using the Web-derived distributions for ODP (ConceptF) leads to a significant reduction over uniform distributions.

Model	SCIENCE		SOCIALSTUDIES	
	CIDE	ODP	CIDE	ODP
ConceptU	7019	5787	13071	9476
ConceptF	n/a	3651	n/a	7244
ConceptL	1461	1060	3479	2432
ConceptLH	1234	1014	2768	2298
ConceptLH+Topics (T=100)	1100	1014	2362	2297

Table 4. Perplexity scores for various models

Learning the word-concept distributions (ConceptL) yields a further significant decrease in perplexity scores compared to the fixed concept distributions as the concepts can now adapt to the corpus. Additionally, accounting for the hierarchy of the concepts (ConceptLH), by propagating words from child concepts to their parents as mentioned before, reduces perplexity even further. If we then add 100 topics to the ConceptLH model (ConceptLH+Topics (T=100) in Table 4), for the CIDE concepts we see another significant reduction in perplexity for both corpora, but no change for the ODP concepts. ODP concepts on their own (ConceptLH models) have lower perplexities than CIDE concepts, so there seems to be more room for improvement with CIDE when topics are added. In addition, ODP has far more concepts (over 10,000) than CIDE (1923), with the result that in the Topics+ODP model less than 1% of the words are assigned to Topics. Overall the ODP concepts produce lower perplexities than CIDE—probably because of the larger number of concepts in ODP, although in general it need not be the case that more concepts lead to better predictions.

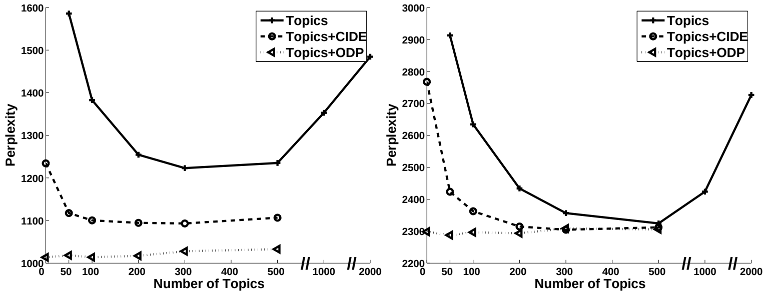


Fig. 6. Comparing perplexity for the Topics model with the ConceptsLH + Topics model on science (left) and social studies (right)

6.3 Varying the Number of Unconstrained Topics

Natural next questions to ask are how would topic models on their own perform and how do the results vary as a function of the number of topics? We address these questions in Figure 6. In this and later experiments in the paper we are using the hierarchical (H) versions of the concept models. The curves in each graph represent topics on their own and topics combined with CIDE and ODP concepts. The x-axis represents the number of topics T used in each model. For example, the point $T = 0$ represents the conceptL models. The results clearly indicate that for any topic model with a fixed number of topics T (a particular point on the x-axis), the performance of the topic model is always improved when concepts are added. The performance improvement is particularly significant on the Science documents, which can be explained by the fact that both CIDE and ODP have well-defined science concepts. It is important to note that the performance difference between topic and concept-topic models is not because of a high number of effective topics ($T + C$) in the concept-topic models. In fact, when we increase the number of topics to $T = 2,000$ for the topic model its perplexity increases significantly possibly due to overfitting. In contrast, the ODP model (for example) is using over 10,000 effective topics ($T + C$) and achieving a lower perplexity score than topics alone. This is a direct illustration of the power of prior knowledge: the constraints represented by human-defined concepts lead to a better language model than what can be obtained with data-driven learning alone.

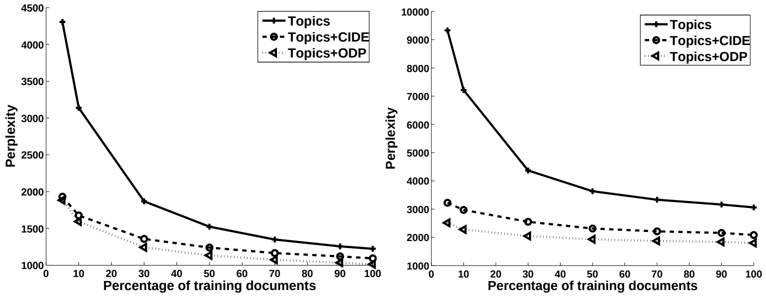


Fig. 7. Perplexity as a function of the amount of training data, testing on science documents, using training data from science (left) and social studies (right)

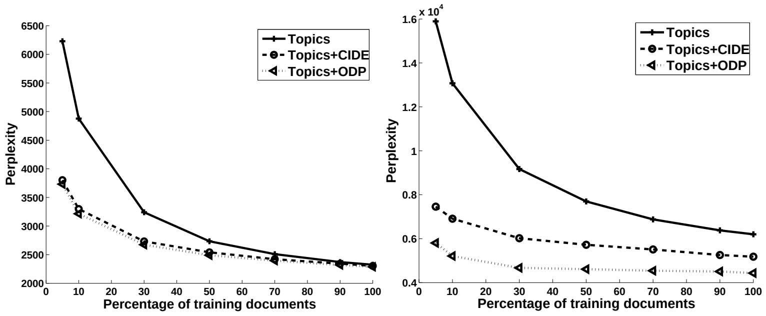


Fig. 8. Perplexity as a function of the amount of training data, testing on social studies documents, using training data from social studies (left) and science (right)

6.4 The Effect of Training Data Size

Finally we look at the effect of varying the amount of training data. The number of topics T used for each model was set to value that produced the lowest perplexity with all of the training data (based on results in Figure 6). Figures 7 and 8 show the perplexity results using science and social studies documents respectively as a test data set. The left plot in each figure shows the results when the training data set and test data set come from the same source and the right plot using different training and test data source.

When there is relatively little training data the concept-topic models have significantly lower perplexity than the topic model. This is a quantitative verification of the oft-quoted idea that "prior knowledge is particularly useful in learning when there is little data." The concept models are helped by the restricted word associations that are manually selected on the basis of their semantic similarity, providing an effective "prior" on words that are expected to co-occur together. The restricted word associations can also help in estimating more accurate word distributions with less data. While it may not be apparent from the figures due the scale used, even at the 100% training data point the concept-topic models have lower perplexity than the topic model (e.g. in Figure 7 at the 100% point on the left, the perplexities of the topic model and the concept-topic model using ODP are 1223.0 and 1013.9 respectively).

As expected, the perplexities are in general higher when a model is trained on one class and predictions are made on a different class (right plots in both the figures). What is notable is that the gap in perplexities between topics and topics+concepts is greater in such cases, i.e., prior knowledge in the form of concepts is even more useful when a model is used on new data that it is different to what it was trained on.

7 Conclusions

We have proposed a general probabilistic text modeling framework that can use both human-defined concepts and data-driven topics. The resulting models allow us to combine the advantages of prior knowledge from the form of ontological concepts and data-driven learning in a systematic manner—for example, the model can automatically place words and documents in a text corpus into a set of human-defined concepts. We also illustrated how concepts can be "tuned" to a corpus to obtain a probabilistic language model leading to improved language models compared with either concepts or topics on their own.

We view the framework presented in this paper as a starting point for exploring a much richer set of models that combine ontological knowledge bases with statistical learning techniques. In Chemudugunta, Smyth and Steyvers [19], we extend the model proposed in this paper to include explicit representation of concept hierarchies. Obvious next steps for exploration are treating concepts and topics differently in the generative model, integrating multiple ontologies and corpora within a single framework, and so forth.

Acknowledgments

The work of the authors was supported in part by the National Science Foundation under Award Number IIS-0083489 as part of the Knowledge Discovery and Dissemination program. In addition, the work of author PS was supported in part by a Google Research Award.

References

1. McGuinness, D.L.: Ontologies come of age. In Fensel, D., Hendler, J.A., Lieberman, H., Wahlster, W., eds.: Spinning the Semantic Web, MIT Press (2003) 171–194
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3** (2003) 993–1022
3. Griffiths, T.L., Steyvers, M.: Finding scientific topics. In: Proc. of Nat'l. Academy of Science. Volume 101. (2004) 5228–5235
4. Handschuh, S., Staab, S., Ciravegna, F.: S-cream — semi-automatic creation of metadata. In: International Conference on Knowledge Engineering and Knowledge Management. (2002)
5. Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., Goranov, M.: Kim - semantic annotation platform. In: International Semantic Web Conference. (2003) 834–849
6. Tang, J., Hong, M., Li, J.Z., Liang, B.: Tree-structured conditional random fields for semantic annotation. In: International Semantic Web Conference. (2006) 640–653
7. Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A., Zien, J.Y.: Semtag and seeker: bootstrapping the semantic web via automated semantic annotation. In: WWW03, ACM (2003) 178–186
8. Hotho, A., Staab, S., Stumme, G.: Text clustering based on background knowledge (technical report 425). Technical report, University of Karlsruhe, Institute AIFB (2003)
9. Gabrilovich, E., Markovitch, S.: Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. J. Mach. Learn. Res. **8** (2007) 2297–2345
10. Ifrim, G., Theobald, M., Weikum, G.: Learning word-to-concept mappings for automatic text classification. In: Proceedings of the 22nd ICML-LWS. (2005) 18–26
11. Boyd-Graber, D., Blei, D., Zhu, X.: A topic model for word sense disambiguation. In: Proc. 2007 Joint Conf. Empirical Methods in Nat'l. Lang. Processing and Compt'l. Nat'l. Lang. Learning. (2007) 1024–1033
12. Brewster, C., Alani, H., Dasmahapatra, S., Wilks, Y.: Data driven ontology evaluation. In: Int'l. Conf. Language Resources and Evaluation. (2004)
13. Alani, H., Brewster, C.: Metrics for ranking ontologies. In: 4th Int'l. EON Workshop, 15th Int'l World Wide Web Conf. (2006)
14. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction and representation of knowledge. Psychological Review **104** (1997) 211–240
15. Griffiths, T.L., Steyvers, M., Tenenbaum, J.B.: Topics in semantic representation. In: Psychological Review. Volume 114. (2007) 211–244
16. Chemudugunta, C., Smyth, P., Steyvers, M.: Modeling general and specific aspects of documents with a probabilistic topic model. In: NIPS 19. (2007) 241–248
17. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database (Language, Speech and Communication). MIT Press (May 1998)
18. Brown, P.F., deSouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Compt'l. Linguistics (1992) 467–479
19. Chemudugunta, C., Smyth, P., Steyvers, M.: Combining concept hierarchies and statistical topic models. In: 17th ACM Conference on Information and Knowledge Management. (2008)

Appendix 1: Inference using Collapsed Gibbs Sampling

Here, we briefly describe the sampling process for the concept-topic model and then describe how sampling for the other models can be viewed as special-cases of this model.

In the concept-topic model, ϕ, ψ and θ correspond to $p(w|t)$ word-topic distributions, $p(w|c)$ word-concept distributions and $p(z|d)$ document level mixtures of topics+concepts respectively. β_ϕ, β_ψ and α correspond to Dirichlet priors on ϕ, ψ and θ multinomial distributions respectively.

In the collapsed Gibbs sampling procedure, the topic assignment variables z_i can be efficiently sampled (after marginalizing the multinomial distributions θ, ϕ and ψ). Point estimates for the marginalized distributions θ, ϕ and ψ can be computed given the assignment labels z_i and predictive distributions are computed by averaging over multiple samples. The sampling equations for the concept-topic model are given by,
case (i): $1 \leq \mathbf{z}_i \leq T$

$$P(\mathbf{z}_i = t|\mathbf{w}_i = w, \mathbf{w}_{-i}, \mathbf{z}_{-i}, \alpha, \beta_\phi) \propto \frac{{C^{WT}}_{wt,-i} + \beta_\phi}{\sum_{w'} {C^{WT}}_{w't,-i} + W\beta_\phi}({C^{(T+C)D}}_{td,-i} + \alpha)$$

case (ii): $\mathbf{z}_i > T$

$$P(\mathbf{z}_i = t|\mathbf{w}_i = w, \mathbf{w}_{-i}, \mathbf{z}_{-i}, \alpha, \beta_\psi) \propto \frac{{C^{WC}}_{wc,-i} + \beta_\psi}{\sum_{w'} {C^{WC}}_{w'c,-i} + N_c\beta_\psi}({C^{(T+C)D}}_{td,-i} + \alpha)$$

where C^{WT}_{wt}, C^{WC}_{wc} are the number of times word w is associated with topic t and concept c respectively, $C^{(T+C)D}_{td}$ is the number of times topic (or concept) t is associated with document d, c = $t - T$ and is only defined for case (ii) and N_c is the number of words associated with concept c. Subscript $-i$ denotes that the word w_i is removed from the counts.

When the concept distributions are fixed (e.g. for the ConceptU model), the inference becomes even simpler as we can just use the fixed distributions in the above equations. Also, note that the topic model and the concept models are special cases of the concept-topic model when C = 0 and T = 0 respectively. Therefore, we can easily adapt the sampling scheme described above to do inference for both these models. It is important to note that the inference for a concept model with N concepts is much faster than the inference of a topic model with N topics. This is because in the case of the concept model we can exploit the sparsity in the word-concept associations — for any word, only the probabilities over concepts that the word is a member of need to be calculated.

We use the standard setup from well-known publications and set α=50/(T+C), β_ϕ=β_ψ=0.01 for models where they are defined. For all our models, we compute the predictive distributions by averaging over 10 different Gibbs chains that are run for 500 iterations and take the last sample to compute the point estimates for the various multinomial distributions.