Combining Concept Hierarchies and Statistical Topic Models

Chaitanya Chemudugunta
Dept. of Computer Science
University of California, Irvine
chandra@ics.uci.edu

Padhraic Smyth
Dept. of Computer Science
University of California, Irvine
smyth@ics.uci.edu

Mark Steyvers
Dept. of Cognitive Science
University of California, Irvine
msteyver@uci.edu

ABSTRACT

Statistical topic models provide a general data-driven framework for automated discovery of high-level knowledge from large collections of text documents. While 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, on the other hand, tend to be semantically richer due to careful selection of words to define concepts but they tend not to cover the themes in a data set exhaustively. In this paper, we propose a probabilistic framework to combine a hierarchy of human-defined semantic concepts with statistical topic models to seek the best of both worlds. Experimental results using two different sources of concept hierarchies and two collections of text documents indicate that this combination leads to systematic improvements in the quality of the associated language models as well as enabling new techniques for inferring and visualizing the semantics of a document.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing–*indexing methods, thesauri*; I.2.6 [Artificial Intelligence]: Learning; I.2.7 [Artificial Intelligence]: Natural Language Processing

General Terms: Algorithms, Experimentation, Human Factors.

Keywords: statistical topic models, unsupervised learning, ontologies, semantic concepts.

1. INTRODUCTION

Latent Dirichlet analysis [2], also referred to as statistical topic modeling [7], is a general framework for automatically summarizing the thematic content of a set of documents. The basic concept underlying statistical topic modeling is that each document is composed of a probability distribution over topics, where each topic is represented as a multinomial probability distribution over words. The document-topic and topic-word distributions are learned automatically from the data in an unsupervised manner. The underlying statistical framework of topic modeling enables a variety of extensions to be developed in a systematic manner (e.g. [10, 1, 9]). An entirely different approach to representing thematic knowledge is to manually define semantic concepts using human knowledge and judgement – this is typically the case with the construction of ontologies and thesauri where a small set of important words are associated with each concept based on prior knowledge. Concept names and sets of relations among concepts (for ontologies) are also often provided.

Concepts (as defined by humans) and topics (as learned from data) represent similar information but in different ways. Human-defined concepts are likely to be more interpretable than topics and can be broader in coverage. Topics on the other hand have the advantage of being tuned to the themes in the particular corpus they are 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. In terms of related work, the models proposed in [8, 3] use topics with prior knowledge for classification and word-sense disambiguation respectively. Chemudugunta et. al. [4] proposed the concept-topic model for combining data-driven topics and semantic concepts to automatically annotate documents. In this paper, we extend the framework in [4] to the hierarchical concept-topic model to take advantage of known hierarchical structure among concepts.

2. HIERARCHICAL MODEL

Concepts are often arranged in a tree-structured hierarchy. Here, we describe the hierarchical concept-topic model (HCTM), that extends the concept-topic model (CTM) in [4] to incorporate the hierarchical structure of the concept set. Similar to the CTM, there are T topics and C concepts in HCTM. For each document d, we introduce a "switch" distribution $p(x|d)$ which determines if a word should be generated via the topic route or the concept route. Every word token in the corpus is associated with a binary switch variable x. If x = 0, the standard topic model (TM) mechanism is used to generate the word. That is, we first select a topic t from a document-specific mixture of topics $p(t|d)$ and generate a word from the word distribution associated with topic t. If x = 1, we generate the word 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. 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 the word 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 — one for each concept.

HCTM represents a document as a weighted combination of mixtures of T topics and C paths through the concept tree:

$$p(w|d) = P(x=0|d)\sum_t p(w|t)p(t|d) + P(x=1|d)\sum_c p(w|c)p(c|d)$$

where $p(c|d) = p(exit|c)p(c|parent(c))...p(.|root)$. HCTM is flexible and can handle any directed-acyclic concept graph. The

Copyright is held by the author/owner(s).
CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

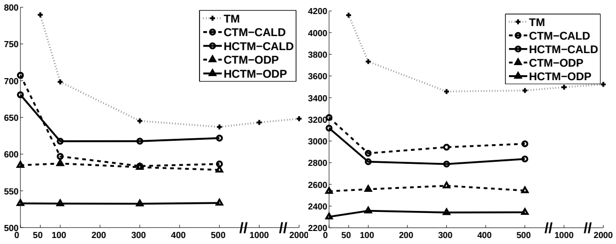

Figure 1: **Comparing perplexity for TM, CTM and HCTM using training documents from science and testing on science (left) and social studies (right) as a function of number of topics**

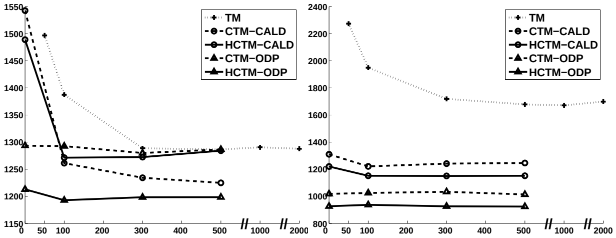

Figure 2: **Comparing perplexity for TM, CTM and HCTM using training documents from social studies and testing on social studies (left) and science (right) as a function of number of topics**

word generation mechanism via the concept route in HCTM is related to the Hierarchical Pachinko Allocation model 2 as described in [9]. There is additional machinery in our model to incorporate T 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. Additional details about the generative process and inference techniques are given in [6].

3. EXPERIMENTS

We use documents from the science and social studies genres of the Touchstone Applied Science Associates (TASA) corpus and concept sets from Cambridge Advanced Learners Dictionary (CALD) and Open Directory Project (ODP) with approximately 2,000 and 10,000 concepts respectively in our experiments. We assess the predictive performance of TM, CTM and HCTM by comparing their perplexity on unseen words in test documents using concepts from CALD and ODP. Perplexity is a quantitative measure to compare language models and is widely used to compare the predictive performance of topic models (e.g. [2, 7, 5]). In the experiments below, we randomly split documents from the science and social studies genres of the TASA corpus into disjoint train and test sets with 90% of the documents included in the train set and the remaining 10% in the test set. For each test document, we use 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.

For the models using concepts, we indicate the concept set used by appending the name of the concept set to the model name, e.g. HCTM-CALD to indicate that HCTM was trained using concepts from the CALD concept set. Figure 1 shows the perplexity of TM, CTM and HCTM using training documents from the science genre in TASA and testing on documents from the science (left) and social studies (right) genres in TASA respectively as a function of number of data-driven topics T. The point $T = 0$ indicates that there are no topics used in the model. The results clearly indicate that incorporating concepts and modeling the concept-hierarchy greatly improves the perplexity of the models. The performance difference is even more significant when the models are trained on one genre of documents and tested on documents from a different genre (e.g. see the right plot of Figure 1), indicating that the models using concepts are robust and can handle noise. TM, on the other hand, is completely data-driven and does not use any human knowledge, so it is not as robust. One important point to note is that this improved performance by the concept models is not due to the high number of effective topics ($T + C$). In fact, even with T = 2,000 topics TM does not improve its perplexity and even shows signs of deterioration in quality in some cases. The corresponding plots for models using training documents from social studies genre in TASA and testing on documents from the social studies (left) and science (right) genres in TASA respectively are shown in Figure 2 with similar qualitative results as in Figure 1. Figures 1 and 2 also allow us to compare the advantages of modeling the hierarchy of the concept sets. In both these figures when $T = 0$, the performance of HCTM is always better than the performance of CTM for all cases and for both concept sets. This effect can be attributed to modeling the correlations of the child concept nodes. More details on the models, the experimental results and the data sets are provided in [6].

4. CONCLUSIONS

We have proposed a probabilistic framework for combining data-driven topics and a hierarchy of semantically-rich human-defined concepts. Experimental results, using two document collections and two concept sets, indicate that using the semantic concepts and modeling the hierarchy of the concept-sets significantly improves the quality of the resulting language models. This improvement is more pronounced when the training documents and test documents belong to different genres. 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, information retrieval, and so forth.

Acknowledgments

The work of the authors was supported in part by NSF Award Number IIS-0083489 as part of the KDD program. In addition the work of author PS was supported in part by a Google Research Award.

5. REFERENCES

[1] D. M. Blei and J. D. Lafferty. Correlated topic models. In *NIPS*, 2005.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[3] D. Boyd-Graber, D. Blei, and X. Zhu. 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*, pages 1024–1033, 2007.

[4] C. Chemudugunta, A. Holloway, P. Smyth, and M. Steyvers. Modeling documents by combining semantic concepts with unsupervised statistical learning. In *International Semantic Web Conference*, 2008.

[5] C. Chemudugunta, P. Smyth, and M. Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS 19*, 2007.

[6] C. Chemudugunta, P. Smyth, and M. Steyvers. Text modeling using unsupervised topic models and concept hierarchies. url: http://arxiv.org/abs/0808.0973. August 2008.

[7] T. L. Griffiths and M. Steyvers. Finding scientific topics. In *Proc. of Nat'l. Academy of Science*, volume 101, pages 5228–5235, 2004.

[8] G. Ifrim, M. Theobald, and G. Weikum. Learning word-to-concept mappings for automatic text classification. In *22nd ICML-LWS*, pages 18–26, 2005.

[9] D. M. Mimno, W. Li, and A. McCallum. Mixtures of hierarchical topics with pachinko allocation. In *ICML*, pages 633–640, 2007.

[10] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. L. Griffiths. Probabilistic author-topic models for information discovery. In *KDD*, pages 306–315, 2004.