Latent Dirichlet Allocation with Topic-in-Set Knowledge*

David Andrzejewski
Computer Sciences Department
University of Wisconsin-Madison
Madison, WI 53706, USA
andrzeje@cs.wisc.edu

Xiaojin Zhu
Computer Sciences Department
University of Wisconsin-Madison
Madison, WI 53706, USA
jerryzhu@cs.wisc.edu

Abstract

Latent Dirichlet Allocation is an unsupervised graphical model which can discover latent topics in unlabeled data. We propose a mechanism for adding partial supervision, called topic-in-set knowledge, to latent topic modeling. This type of supervision can be used to encourage the recovery of topics which are more relevant to user modeling goals than the topics which would be recovered otherwise. Preliminary experiments on text datasets are presented to demonstrate the potential effectiveness of this method.

1 Introduction

Latent topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have emerged as a useful family of graphical models with many interesting applications in natural language processing. One of the key virtues of LDA is its status as a fully generative probabilistic model, allowing principled extensions and variations capable of expressing rich problem domain structure (Newman et al., 2007; Rosen-Zvi et al., 2004; Boyd-Graber et al., 2007; Griffiths et al., 2005).

LDA is an unsupervised learning model. This work aims to add supervised information in the form of latent topic assignments to LDA. Traditionally, topic assignments have been denoted by the variable z in LDA, and we will call such supervised information "z-labels." In particular, a z-label is the knowledge that the topic assignment for a given word position is within a subset of topics. As such, this work is a combination of unsupervised model and supervised knowledge, and falls into the category similar to constrained clustering (Basu et al., 2008) and semi-supervised dimensionality reduction (Yang et al., 2006).

1.1 Related Work

A similar but simpler type of topic labeling information has been applied to computer vision tasks. Topic modeling approaches have been applied to scene modeling (Sudderth et al., 2005), segmentation, and classification or detection (Wang and Grimson, 2008). In some of these vision applications, the latent topics themselves are assumed to correspond to object labels. If labeled data is available, either all (Wang and Mori, 2009) or some (Cao and Fei-Fei, 2007) of the z values can be treated as observed, rather than latent, variables. Our model extends z-labels from single values to subsets, thus offer additional model expressiveness.

If the topic-based representations of documents are to be used for document clustering or classification, providing z-labels for words can be seen as similar to semi-supervised learning with labeled features (Druck et al., 2008). Here the words are features, and z-label guidance acts as a feature label. This differs from other supervised LDA variants (Blei and McAuliffe, 2008; Lacoste-Julien et al., 2008) which use document label information.

The ΔLDA model for statistical software debugging (Andrzejewski et al., 2007) partitions the topics into 2 sets: "usage" topics which can appear in all

* We would like to acknowledge the assistance of Brandi Gancarz with the biological annotations. This work is supported in part by the Wisconsin Alumni Research Foundation.

documents, and "bug" topics which can only appear in a special subset of documents. This effect was achieved by using different α hyperparameters for the 2 subsets of documents. z-labels can achieve the same effect by restricting the z's in documents outside the special subset, so that the z's cannot assume the "bug" topic values. Therefore, the present approach can be viewed as a generalization of ΔLDA.

Another perspective is that our z-labels may guide the topic model towards the discovery of secondary or non-dominant statistical patterns in the data (Chechik and Tishby, 2002). These topics may be more interesting or relevant to the goals of the user, but standard LDA would ignore them in favor of more prominent (and perhaps orthogonal) structure.

2 Our Model

2.1 Review of Latent Dirichlet Allocation

We briefly review LDA, following the notation of (Griffiths and Steyvers, 2004) [1]. Let there be T topics. Let $\mathbf{w} = w_1 \dots w_n$ represent a corpus of D documents, with a total of n words. We use d_i to denote the document of word w_i, and z_i the hidden topic from which w_i is generated. Let $\phi_j^{(w)} = p(w|z = j)$, and $\theta_j^{(d)} = p(z = j)$ for document d. LDA involves the following generative model:

$$\begin{aligned} \theta &\sim \text{Dirichlet}(\alpha) && (1)\\ z_i|\theta^{(d_i)} &\sim \text{Multinomial}(\theta^{(d_i)}) && (2)\\ \phi &\sim \text{Dirichlet}(\beta) && (3)\\ w_i|z_i,\phi &\sim \text{Multinomial}(\phi_{z_i}), && (4) \end{aligned}$$

where α and β are hyperparameters for the document-topic and topic-word Dirichlet distributions, respectively. Even though they can be vector valued, for simplicity we assume α and β are scalars, resulting in symmetric Dirichlet priors.

Given our observed words $\mathbf{w}$, the key task is inference of the hidden topics $\mathbf{z}$. Unfortunately, this posterior is intractable and we resort to a Markov Chain Monte Carlo (MCMC) sampling scheme, specifically Collapsed Gibbs Sampling (Griffiths and Steyvers, 2004). The full conditional equation used for sampling individual z_i values from the posterior is given by

$$P(z_i = v|\mathbf{z}_{-i}, \mathbf{w}, \alpha, \beta) \propto \left(\frac{n_{-i,v}^{(d)} + \alpha}{\sum_u^T (n_{-i,u}^{(d)} + \alpha)}\right)\left(\frac{n_{-i,v}^{(w_i)} + \beta}{\sum_{w'}^W (\beta + n_{-i,v}^{(w')})}\right) \quad (5)$$

where $n_{-i,v}^{(d)}$ is the number of times topic v is used in document d, and $n_{-i,v}^{(w_i)}$ is the number of times word w_i is generated by topic v. The $-i$ notation signifies that the counts are taken omitting the value of z_i.

2.2 Topic-in-Set Knowledge: z-labels

Let

$$q_{iv} = \left(\frac{n_{-i,v}^{(d)} + \alpha}{\sum_u^T (n_{-i,u}^{(d)} + \alpha)}\right)\left(\frac{n_{-i,v}^{(w_i)} + \beta}{\sum_{w'}^W (\beta + n_{-i,v}^{(w')})}\right).$$

We now define our z-labels. Let $C^{(i)}$ be the set of possible z-labels for latent topic z_i. We set a hard constraint by modifying the Gibbs sampling equation with an indicator function $\delta(v \in C^{(i)})$, which takes on value 1 if $v \in C^{(i)}$ and is 0 otherwise:

$$P(z_i = v|\mathbf{z}_{-i}, \mathbf{w}, \alpha, \beta) \propto q_{iv}\delta(v \in C^{(i)}) \quad (6)$$

If we wish to restrict z_i to a single value (e.g., $z_i = 5$), this can now be accomplished by setting $C^{(i)} = \{5\}$. Likewise, we can restrict z_i to a subset of values $\{1, 2, 3\}$ by setting $C^{(i)} = \{1, 2, 3\}$. Finally, for unconstrained z_i we simply set $C^{(i)} = \{1, 2, ..., T\}$, in which case our modified sampling (6) reduces to the standard Gibbs sampling (5).

This formulation gives us a flexible method for inserting prior domain knowledge into the inference of latent topics. We can set $C^{(i)}$ independently for every single word w_i in the corpus. This allows us, for example, to force two occurrences of the same word (e.g., "Apple pie" and "Apple iPod") to be explained by different topics. This effect would be impossible to achieve by using topic-specific asymmetric β vectors and setting some entries to zero.

This hard constraint model can be relaxed. Let $0 \leq \eta \leq 1$ be the strength of our constraint, where $\eta = 1$ recovers the hard constraint (6) and $\eta = 0$ recovers unconstrained sampling (5):

$$P(z_i = v|\mathbf{z}_{-i}, \mathbf{w}, \alpha, \beta) \propto q_{iv}\left(\eta\delta(v \in C^{(i)}) + 1 - \eta\right).$$

[1] We enclose superscripts in parentheses in this paper.

While we present the z-label constraints as a mechanical modification to the Gibbs sampling equations, it can be derived from an undirected extension of LDA (omitted here) which encodes z-labels. The soft constraint Gibbs sampling equation arises naturally from this formulation, which is the basis for the First-Order Logic constraints described later in the future work section.

3 Experiments

We now present preliminary experimental results to demonstrate some interesting applications for topic-in-set knowledge. Unless otherwise specified, symmetric hyperparameters $\alpha = .5$ and $\beta = .1$ were used and all MCMC chains were run for 2000 samples before estimating ϕ and θ from the final sample, as in (Griffiths and Steyvers, 2004).

3.1 Concept Expansion

We explore the use of topic-in-set for identifying words related to a target concept, given a set of seed words associated with that concept. For example, a biological expert may be interested in the concept "translation". The expert would then provide a set of seed words which are strongly related to this concept, here we assume the seed word set {translation,trna,anticodon,ribosome}. We add the hard constraint that $z_i = 0$ for all occurrences of these four words in our corpus of approximately 9,000 yeast-related abstracts.

We ran LDA with the number of topics $T = 100$, both with and without the z-label knowledge on the seed words. Table 1 shows the most probable words in selected topics from both runs. Table 1a shows Topic 0 from the constrained run, while Table 1b shows the topics which contained seed words among the top 50 most probable words from the unconstrained run.

In order to better understand the results, these top words were annotated for relevance to the target concept (translation) by an outside biological expert. The words in Table 1 were then colored blue if they were one of the original seed words, red if they were judged as relevant, and left black otherwise. From a quick glance, we can see that Topic 0 from the constrained run contains more relevant terms than Topic 43 from the standard LDA run. Topic 31 has a similar number of relevant terms, but taken together we can see that the emphasis of Topic 31 is slightly off-target, more focused on "mRNA turnover" than "translation". Likewise, Topic 73 seems more focused on the ribosome itself than the process of translation. Overall, these results demonstrate the potential effectiveness of z-label information for guiding topic models towards a user-seeded concept.

3.2 Concept Exploration

Suppose that a user has chosen a set of terms and wishes to discover different topics related to these terms. By constraining these terms to only appear in a restricted set of topics, these terms will be *concentrated* in the set of topics. The split within those set of topics may be different from what a standard LDA will produce, thus revealing new information within the data.

To make this concrete, say we are interested in the location "United Kingdom". We seed this concept with the following LOCATION-tagged terms {britain, british, england, uk, u.k., wales, scotland, london}. These terms are then restricted to appear only in the first 3 topics. Our corpus is an entity-tagged Reuters newswire corpus used for the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003). In order to focus on our target location, we also restrict all other LOCATION-tagged tokens to *not* appear in the first 3 topics. For this experiment we set $T = 12$, arrived at by trial-and-error in the baseline (standard LDA) case.

The 50 most probable words for each topic are shown in Figure 2, and tagged entities are prefixed with their tags for easy identification. Table 2a shows the top words for the first 3 topics of our z-label run. These three topics are all related to the target LOCATION United Kingdom, but they also split nicely into business, cricket, and soccer. Words which are highly relevant to each of these 3 concepts are colored blue, red, and green, respectively.

In contrast, in Table 2b we show topics from standard LDA which contain any of the "United Kingdom" LOCATION terms (which are underlined) among the 50 most probable words for that topic. We make several observations about these topics. First, standard LDA Topic 0 is mostly concerned with political unrest in Russia, which is not particu-

Topic 0	translation, ribosomal, trna, rrna, initiation, ribosome, protein, ribosomes, is, factor, processing, translational nucleolar, pre-rrna, synthesis, small, 60s, eukaryotic, biogenesis, subunit, trnas, subunits, large, nucleolus factors, 40, synthetase, free, modification, rna, depletion, eif-2, initiator, 40s, ef-3, anticodon, maturation 18s, eif2, mature, eif4e, associated, synthetases, aminoacylation, snornas, assembly, eif4g, elongation

(a) Topic 0 with z-label

Topic 31	mrna, translation, initiation, mrnas, rna, transcripts, 3, transcript, polya, factor, 5, translational, decay, codon decapping, factors, degradation, end, termination, eukaryotic, polyadenylation, cap, required, efficiency synthesis, show, codons, abundance, rnas, aug, nmd, messenger, turnover, rna-binding, processing, eif2, eif4e eif4g, cf, occurs, pab1p, cleavage, eif5, cerevisiae, major, primary, rapid, tail, efficient, upf1p, eif-2
Topic 43	type, is, wild, yeast, trna, synthetase, both, methionine, synthetases, class, trnas, enzyme, whereas, cytoplasmic because, direct, efficiency, presence, modification, aminoacylation, anticodon, either, eukaryotic, between different, specific, discussed, results, similar, some, met, compared, aminoacyl-trna, able, initiator, sam not, free, however, recognition, several, arc1p, fully, same, forms, leads, identical, responsible, found, only, well
Topic 73	ribosomal, rrna, protein, is, processing, ribosome, ribosomes, rna, nucleolar, pre-rrna, rnase, small, biogenesis depletion, subunits, 60s, subunit, large, synthesis, maturation, nucleolus, associated, essential, assembly components, translation, involved, rnas, found, component, mature, rp, 40s, accumulation, 18s, 40, particles snornas, factors, precursor, during, primary, rrnas, 35s, has, 21s, specifically, results, ribonucleoprotein, early

(b) Standard LDA Topics

Figure 1: Concept seed words are colored blue, other words judged relevant to the target concept are colored red.

larly related to the target location. Second, Topic 2 is similar to our previous business topic, but with a more US-oriented slant. Note that "dollar" appears with high probability in standard LDA Topic 2, but not in our z-label LDA Topic 0. Standard LDA Topic 8 appears to be a mix of both soccer and cricket words. Therefore, it seems that our topic-in-set knowledge helps in distilling topics related to the seed words.

Given this promising result, we attempted to repeat this experiment with some other nations (United States, Germany, China), but without much success. When we tried to restrict these LOCATION words to the first few topics, these topics tended to be used to explain other concepts unrelated to the target location (often other sports). We are investigating the possible causes of this problem.

4 Conclusions and Future Work

We have defined Topic-in-Set knowledge and demonstrated its use within LDA. As shown in the experiments, the partial supervision provided by z-labels can encourage LDA to recover topics relevant to user interests. This approach combines the pattern-discovery power of LDA with user-provided guidance, which we believe will be very attractive to practical users of topic modeling.

Future work will deal with at least two important issues. First, when will this form of partial supervision be most effective or appropriate? Our experimental results suggest that this approach will struggle if the user's target concepts are simply not prevalent in the text. Second, can we modify this approach to express richer forms of partial supervision? More sophisticated forms of knowledge may allow users to specify their preferences or prior knowledge more effectively. Towards this end, we are investigating the use of First-Order Logic in specifying prior knowledge. Note that the set z-labels presented here can be expressed as simple logical formulas. Extending our model to general logical formulas would allow the expression of more powerful relational preferences.

References

David Andrzejewski, Anne Mulhern, Ben Liblit, and Xiaojin Zhu. 2007. Statistical debugging using latent topic models. In Stan Matwin and Dunja Mladenic, editors, *18th European Conference on Machine Learning*, Warsaw, Poland.

Topic 0	million, company, 's, year, shares, net, profit, half, group, [I-ORG]corp, market, sales, share, percent expected, business, loss, stock, results, forecast, companies, deal, earnings, statement, price, [I-LOC]london billion, [I-ORG]newsroom, industry, newsroom, pay, pct, analysts, issue, services, analyst, profits, sale added, firm, [I-ORG]london, chief, quarter, investors, contract, note, tax, financial, months, costs
Topic 1	[I-LOC]england, [I-LOC]london, [I-LOC]britain, cricket, [I-PER]m., overs, test, wickets, scores, [I-PER]ahmed [I-PER]paul, [I-PER]wasim, innings, [I-PER]a., [I-PER]akram, [I-PER]mushtaq, day, one-day, [I-PER]mark, final [I-LOC]scotland, [I-PER]waqar, [I-MISC]series, [I-PER]croft, [I-PER]david, [I-PER]younis, match, [I-PER]ian total, [I-MISC]english, [I-PER]khan, [I-PER]mullally, bat, declared, fall, [I-PER]d., [I-PER]g., [I-PER]j. bowling, [I-PER]r., [I-PER]robert, [I-PER]s., [I-PER]steve, [I-PER]c. captain, golf, tour, [I-PER]sohail, extras [I-ORG]surrey
Topic 2	soccer, division, results, played, standings, league, matches, halftime, goals, attendance, points, won, [I-ORG]st drawn, saturday, [I-MISC]english, lost, premier, [I-MISC]french, result, scorers, [I-MISC]dutch, [I-ORG]united [I-MISC]scottish, sunday, match, [I-LOC]london, [I-ORG]psv, tabulate, [I-ORG]hapoel, [I-ORG]sydney, friday summary, [I-ORG]ajax, [I-ORG]manchester, tabulated, [I-MISC]german, [I-ORG]munich, [I-ORG]city [I-MISC]european, [I-ORG]rangers, summaries, weekend, [I-ORG]fc, [I-ORG]sheffield, wednesday, [I-ORG]borussia [I-ORG]fortuna, [I-ORG]paris, tuesday

(a) Topics with set z-labels

Topic 0	police, 's, people, killed, [I-MISC]russian, friday, spokesman, [I-LOC]moscow, told, rebels, group, officials [I-PER]yeltsin, arrested, found, miles, km, [I-PER]lebed, capital, thursday, tuesday, [I-LOC]chechnya, news saturday, town, authorities, airport, man, government, state, agency, plane, reported, security, forces city, monday, air, quoted, students, region, area, local, [I-LOC]russia, [I-ORG]reuters, military, [I-LOC]london held, southern, died
Topic 2	percent, 's, market, thursday, july, tonnes, week, year, lower, [I-LOC]u.s., rate, prices, billion, cents, dollar friday, trade, bank, closed, trading, higher, close, oil, bond, fell, markets, index, points, rose demand, june, rates, september, traders, [I-ORG]newsroom, day, bonds, million, price, shares, budget, government growth, interest, monday, [I-LOC]london, economic, august, expected, rise
Topic 5	's, match, team, win, play, season, [I-MISC]french, lead, home, year, players, [I-MISC]cup, back, minutes champion, victory, time, n't, game, saturday, title, side, set, made, wednesday, [I-LOC]england league, run, club, top, good, final, scored, coach, shot, world, left, [I-MISC]american, captain [I-MISC]world, goal, start, won, champions, round, winner, end, years, defeat, lost
Topic 8	division, [I-LOC]england, soccer, results, [I-LOC]london, [I-LOC]pakistan, [I-MISC]english, matches, played standings, league, points, [I-ORG]st, cricket, saturday, [I-PER]ahmed, won, [I-ORG]united, goals [I-PER]wasim, [I-PER]akram, [I-PER]m., [I-MISC]scottish, [I-PER]mushtaq, drawn, innings, premier, lost [I-PER]waqar, test, [I-PER]croft, [I-PER]a., [I-PER]younis, declared, wickets, [I-ORG]hapoel, [I-PER]mullally [I-ORG]sydney, day, [I-ORG]manchester, [I-PER]khan, final, scores, [I-PER]d., [I-MISC]german, [I-ORG]munich [I-PER]sohail, friday, total, [I-LOC]oval
Topic 10	[I-LOC]germany, 's, [I-LOC]italy, [I-LOC]u.s., metres, seconds, [I-LOC]france, [I-LOC]britain, [I-LOC]russia world, race, leading, [I-LOC]sweden, [I-LOC]australia, [I-LOC]spain, women, [I-MISC]world, [I-LOC]belgium [I-LOC]netherlands, [I-PER]paul, [I-LOC]japan, [I-MISC]olympic, [I-LOC]austria, [I-LOC]kenya, men, time results, [I-LOC]brussels, [I-MISC]cup, [I-LOC]canada, final, minutes, record, [I-PER]michael, meeting, round [I-LOC]norway, friday, scores, [I-PER]mark, [I-PER]van, [I-LOC]ireland, [I-PER]peter, [I-MISC]grand [I-MISC]prix, points, saturday, [I-LOC]finland, cycling, [I-ORG]honda

(b) Standard LDA Topics

Figure 2: Topics containing "United Kingdom" location words. Words related to business are colored blue, cricket red, and soccer green.

Sugato Basu, Ian Davidson, and Kiri Wagstaff, editors. 2008. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC Press.

David Blei and Jon McAuliffe. 2008. Supervised topic models. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. MIT Press,

Cambridge, MA.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1024–1033.

Liangliang Cao and Li Fei-Fei. 2007. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV*, pages 1–8.

Gal Chechik and Naftali Tishby. 2002. Extracting relevant structures with side information. In *NIPS 15*, pages 857–864. MIT press.

Gregory Druck, Gideon Mann, and Andrew McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *SIGIR 2008*, pages 595–602, New York, NY, USA. ACM.

Thomas Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl. 1):5228–5235.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2005. Integrating topics and syntax. In *NIPS 17*.

S. Lacoste-Julien, F. Sha, and M. Jordan. 2008. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems 21 (NIPS08)*.

David Newman, Kat Hagedorn, Chaitanya Chemudugunta, and Padhraic Smyth. 2007. Subject metadata enrichment using statistical topic models. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 366–375, New York, NY, USA. ACM.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence (UAI)*, pages 487–494, Arlington, Virginia, United States. AUAI Press.

Erik B. Sudderth, Antonio B. Torralba, William T. Freeman, and Alan S. Willsky. 2005. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, pages 1331–1338.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, pages 142–147, Edmonton, Canada.

Xiaogang Wang and Eric Grimson. 2008. Spatial latent dirichlet allocation. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *NIPS 20*, pages 1577–1584. MIT Press, Cambridge, MA.

Yang Wang and Greg Mori. 2009. Human action recognition by semi-latent topic models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xin Yang, Haoying Fu, Hongyuan Zha, and Jesse Barlow. 2006. Semi-supervised nonlinear dimensionality reduction. In *ICML-06, 23nd International Conference on Machine Learning*.