From Topic Models to Semi-supervised Learning: Biasing Mixed-Membership Models to Exploit Topic-Indicative Features in Entity Clustering

Ramnath Balasubramanyan[1], Bhavana Dalvi[1], and William W. Cohen[1,2]

[1] Language Technologies Institute, Carnegie Mellon University
{rbalasub,bbd,wcohen}@cs.cmu.edu
[2] Machine Learning Department, Carnegie Mellon University

Abstract. We present methods to introduce different forms of supervision into mixed-membership latent variable models. Firstly, we introduce a technique to bias the models to exploit *topic-indicative* features, i.e. features which are *apriori* known to be good indicators of the latent topics that generated them. Next, we present methods to modify the Gibbs sampler used for approximate inference in such models to permit injection of stronger forms of supervision in the form of labels for features and documents, along with a description of the corresponding change in the underlying generative process. This ability allows us to span the range from unsupervised topic models to semi-supervised learning in the same mixed membership model. Experimental results from an entity-clustering task demonstrate that the biasing technique and the introduction of feature and document labels provide a significant increase in clustering performance over baseline mixed-membership methods.

1 Introduction

Topic modeling based on Latent Dirichlet Allocation (LDA) [6] has become a popular tool for data exploration, dimensionality reduction and for facilitating myriad other tasks [2,1,12]. As a fully unsupervised technique, however, topic models are unequipped to utilize limited supervisory information, e.g. feature labels and document cluster membership. In this paper, we introduce methods to incorporate progressively stronger forms of weak supervision to influence the formation of topics that respect information that we might have about the latent structure.

First, we present a method to bias mixed-membership models (such as topic models) to better exploit known *topic-indicative* features. Unsupervised topic models do not necessarily optimally utilize topic-indicative features, i.e. features that are known to be strongly indicative of the latent topics of the documents. The biasing towards topic-indicative features serves to control the latent role distribution of the features, i.e., the degree of *polysemy*, and its strength can be adjusted to control the degree of polysemy permitted.

The flexibility of the biased models is examined by using it to cluster entities found in HTML pages [9]. While our model can be used for a variety of tasks, we focus on the HTML entities clustering tasks since it requires the use of several kinds of features (obtained from semi-structured data from the tables) and permits us to demonstrate ways

H. Blockeel et al. (Eds.): ECML PKDD 2013, Part II, LNAI 8189, pp. 628–642, 2013.

in which intuition and limited supervision about different kinds of features can be incorporated. In this task, potentially useful features of an entity includes features like the headers of columns (e.g. the entity *apple* might be found under the headers *company* or *fruit*) and domains (e.g. food.com, finance.com, etc.). The biasing technique presented could be used to capture domain knowledge that features of a certain type are more topic-indicative than other features. When the bias term is set high, the features to which it is applied are deemed to be more strongly indicative of topic and are strongly discouraged from assuming multiple latent roles in the mixed membership model. The bias is accomplished using a regularization term in the model which represents a noisy copy of the entropy of the latent role distribution of the word. The polysemy is reduced by pushing the entropy towards a pre-specified desired value that is a hyperparameter to the model.

Next, we show that stronger forms of supervision to the model in the form of feature and document labels can be injected into the model to achieve modeling flexibility to obtain models that range from fully unsupervised topic models to semi-supervised models. This form of light supervision can be in the form of known latent roles for certain subsets of features or known latent roles for documents which exhibit very slight mixed-membership characteristics. The supervision is incorporated into the model by modifying the Gibbs sampling procedure used for approximate inference.

The rest of the paper is organized as follows. Section 2 describes the mixed-membership latent variable model based approach to the entity clustering task. Next, we describe the biasing technique to exploit topic-indicative features in Section 3 and the approach to incorporate feature and document labels in Section 4. Experimental results are presented in Section 5. Finally, we present a short survey of related work in Section 6, followed by the conclusion.

2 Entity Clustering

Latent-variable mixed-membership models based on LDA have been used for a variety of tasks in NLP. Here, we use it for the task of clustering entities that are extracted from tables in HTML documents crawled from the web. Dalvi et al. [9] describe the task in detail.

In this task, the dataset consists of tables of entities extracted from HTML pages. For instance, it could contain a table of companies, tables of American football teams, etc. The goal of the task is to cluster entities of the same semantic class together. Therefore, if the dataset includes a table of fruits with apples, grapes and oranges, and another table with oranges, peaches and bananas, the goal of the task is to recover a cluster of fruits which includes apples, grapes, oranges, peaches and bananas.

Surface terms in such HTML tables frequently have multiple senses. For example, consider the term *apple*, which is found in tables of companies and fruits among others. Therefore we require a model that is capable of distinguishing the sense of the term to prevent companies and fruits from being collapsed into one cluster based on the term *apple* co-occurring with both companies and fruits. Mixed-membership models can account for the multiple-sense problem by assigning partial membership in both clusters to the entity. Typically, entity clustering has been based on distributional similarity based approaches or by using Hearst patterns [13]. In this task however, since we

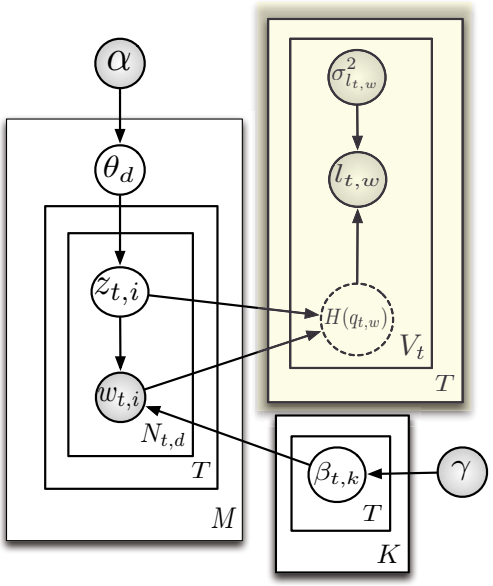

Fig. 1. Biased Link-LDA model to Exploit Topic Indicative Features

are dealing with entities in HTML tables as opposed to entity mentions in free text, we use a different set of features to assist in the clustering, namely:

a) co-occurring entities,
b) co-occuring entity pairs that the entity is observed with,
c) the tableid-columnid combinations under which the entity was observed,
d) web domains in which the entity was observed,
e) the hyponyms that are associated with an entity (extracted using Hearst patterns).

This task can therefore be seen as distributional clustering with a different set of contextual features than the free text features usually used. For every unique entry found in the collection of tables in a dataset, we construct a *"document"* in the LDA sense with the above five kinds of *"words"*. The document is represented by a set of bags of words, one for each kind of feature used. A document for the entity *apple*, for example might consists of the following bags -

a) co-occurring entities {*orange*, *apple*, *microsoft*, . . . },
b) entity pairs {*orange:apple*, *google:apple* . . . },
c) column ids {tab:326::colid::1 . . .},
d) domains {business.com, produce.com . . .},
e) hyponyms {*stocks*, *juice*, *tech companies* . . .}.

These different classes of features are modeled using the Link-LDA model [10]. Figure 1 shows the plate diagram of the graphical model. The variables that are under the yellow shaded rectangle provide the bias that is introduced in later sections, and are not part of the regular Link-LDA model. In the generative story for the model, a document has T kinds of "words". For instance, in a corpus of academic papers, the kinds of words could be author names, words in the abstract, words in the body, references to other papers etc. For every document in the corpus of size M, a distribution over K topics θ is first drawn. Then the words of all T kinds are drawn by first sampling a topic indicator k for the word from θ and then drawing the word from the per-type topic word distributions $\beta_{t,k}$. Since exact inference is intractable for the model, we use a collapsed Gibbs sampler [19] for approximate inference. θ, the document topic distribution obtained after inference provides an estimate for the predicted cluster membership of an entity document.

The predicted clusters are evaluated using *Normalized Mutual Information (NMI)*. This information theory based score measures the amount of information about the true clusters that is encoded by the predicted topic/cluster distributions. NMI can be used in mixed-membership scenarios since the true cluster distribution and predicted topic distribution can have probability mass in more than one cluster. Additionally, the number of true clusters and topics do not have to be the same and therefore no mapping from topics to clusters is required. To compute NMI between the true cluster label distribution and predicted distributions for the test entity set, we first compute Ω the predicted distribution of topics which is equal to $\frac{\sum_{e \in \text{test set}} \theta_e}{|\text{test set}|}$. Let $\mathcal{C}$ be the distribution over true cluster labels, then NMI is defined as $\frac{I(\Omega;\mathcal{C})}{(H(\Omega)+H(\mathcal{C}))/2}$, where I indicates mutual information. It should be noted that while the model returns mixed-membership assignments for entities, the human labeling scheme that was used provides only one true cluster assignment for an entity. We however present a qualitative analysis of the advantages of mixed-membership modeling in Section 5.

Entity clustering experiments were performed using the WebSets datasets [9], namely — the Asia NELL, Clueweb Sports, CSEAL Useful, Delicious Music, Delicious Sports and Toy Apple datasets. The Asia NELL dataset was collected using the ASIA system [24] using hypernyms of NELL [7] entities as queries. The Clueweb Sports dataset consists of tables extracted from Sports related pages in the Clueweb dataset. The Delicious music and sports datasets consist of tables from subsets of the DAI-Labor [25] Delicious corpus that were tagged as music and sports respectively. The Toy Apple dataset is a small toy dataset constructed using the SEAL [8] system to create set-expansion lists using the query "Apple", which is a typical example of a multi-sense entity (as a fruit and as a company). It is used primarily to illustrate the effects of clustering mixed membership entities. Statistics about the datasets are shown in Table 1.

In the WebSets approach by Dalvi et al., triplets of entities from HTML tables are extracted and then clustered. Their approach also proposes a method to propose labels for the clusters. It should be noted that their approach clusters triples of entities rather than individual entities which makes it hard to directly compare performance with the method proposed in this paper.

Table 1. Dataset Statistics

		Size of vocabulary				
Dataset	entities	co-occurring entities	entity pairs	column ids	domains	hyponyms
Asia NELL	33455	18309	141352	9477	3207	31833
Clueweb Sports	29113	28891	354614	59117	8088	28618
CSEAL Useful	34565	24340	217328	7337	2118	28381
Delicious Music	18074	9748	106401	7564	1633	24934
Delicious Sports	6786	3183	24147	2050	509	16380
Toy Apple	2411	423	4737	109	53	2826

3 Biasing Topic Indicative Features Using Entropic Regularization

One of the attractive attributes of topic models is that they require no supervision in terms of data annotation. However, in many situations, limited amounts of labeled data may be available. We present an approach to bias topic models to utilize weak knowledge about features. Specifically, we aim to make the model exploit *topic indicative features*, which are a subset of features that are known beforehand to be strongly indicative of topic. For instance in the toy apple example, co-occurring entities of the ambiguous entity *apple* are topic indicative. Co-occurring entities such as *Google* and *Microsoft* are indicative of the company topic where as co-occurring entities like *grape* and *banana* indicate the fruit topic. The bias is introduced into the model via a regularization term that constrains the freedom of specific features to take on multiple latent roles.

The LDA model and its extensions allow the same word to belong to different topics when they are instantiated multiple times in the corpus. This freedom is essential in modeling *polysemy*. While this freedom is useful, we aim to control this freedom for features that are topic-indicative. Following the idea illustrated in Figure 1, we present a entropy based regularization technique based on pseudo-observed variables [4], which directly controls the freedom of words to take on different latent topics, by penalizing high entropies in their topic distributions. It should be noted that sparsity in a document's topic membership vector can be achieved using sparse priors, but sparsity in a words' latent role distribution cannot be similarly obtained since these distributions are not explicitly sampled in a topic model. The addition of the regularization term however allows us to impose such preferences by relaxing the conditional independence between topic multinomials in LDA-like models.

Let n_{tkw} be the number of times a word w of type t was observed with latent role k. The topic distribution of a word w of kind t in a topic model can be defined as $q^{(k)}_{t,w} = \frac{n_{tkw}}{\sum_{k'} n_{tk'w}}, k \in 1, \ldots, K$. $q_{t,w}$ therefore shows the degree of polysemy exhibited by a word in the model. The Shannon entropy of this distribution is denoted by $H(q_{t,w})$.

We now introduce word topic distribution entropy regularization by adding pseudo-observed variables, $l_{t,w}$ (Figure 1), one for each word of every kind t in the vocabulary V_t, which are noisy copies of $H(q_{t,w})$. These noisy copies are drawn from a one-sided

truncated Gaussian, whose mass lies only on values between 0 and $\log_2 K$, with mean $H(q_{t,w})$ and variance $\sigma^2_{l_{t,w}}$, which is a hyperparameter to the model. The density function is given by

$$p(l_{t,w}|h, \sigma^2_{l_{t,w}}) = \begin{cases} \frac{1}{C} \exp\left(\frac{-(h-l_{t,w})^2}{2\sigma^2_{l_{t,w}}}\right) & \text{for } 0 \leq l_{t,w} \leq \log_2 K \\ 0, & \text{otherwise.} \end{cases}$$

$C = \int_{h'=0}^{\log_2 K} \exp\left(\frac{-(h^{'} - l_{t,w})^2}{2\sigma^2_{l_{t,w}}}\right) dh.$

The joint distribution of the model with regularization is defined as:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\alpha, \gamma, \mathbf{l_{t,w}}, \boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{l_{t,w}}}) = \prod_{d=1}^{M} \text{Dir}(\theta_d|\alpha)$$
$$\left(\prod_{t=1}^{T}\left(\prod_{i=1}^{N_{t,d}} \theta_d^{z_{t,i}} \beta_{t,z_{t,i}}^{(w_{t,i})}\right)\right) \prod_t \prod_k \text{Dir}(\beta_{t,k}|\gamma) \prod_t \prod_{w \in V_t} \exp \frac{-\left(l_{t,w} - H(q_{t,w})\right)^2}{2\sigma^2_{l_{t,w}}} / C \tag{1}$$

Approximate inference in the model is performed using a collapsed Gibbs sampler. Let n_{dk} be the number of words in document d that were assigned to topic k. The equation to sample a topic indicator for a word $w_{t,i}$ i.e. the i-th word of type t in d, is given by

$$p(z_{t,i} = k|\mathbf{l_{t,w}}, w_{t,i}, \mathbf{z}^{\neg \mathbf{t,i}}, \mathbf{w}^{\neg \mathbf{t,i}}, \alpha, \gamma, \boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{l_{t,w}}}) \propto$$
$$(n_{dk}^{\neg t,i} + \alpha) \frac{n_{tkw_{t,i}}^{\neg t,i} + \gamma}{\sum_{w'} n_{tkw'}^{\neg i} + |V_t|\gamma} \times \exp\left(\frac{-(H(q_{t,w_{t,i}}) - l_{t,w_{t,i}})^2}{2\sigma^2_{t,l_{t,w}}}\right) \tag{2}$$

During the Gibbs sampling process, the inference procedure tends to push the mean of the Gaussians i.e. $H(q_{t,w})$ close to the preset $l_{t,w}$ values. For topic-indicative features, we set $l_{t,w}$ to 0 which penalizes large entropies in the topic distributions of such features, therefore driving the inference procedure to return low entropy models. $\sigma^2_{l_w}$ dictates the strictness of the penalty.

It should be noted that an alternate method to achieve sparsity is to modify the priors. Replacing the Dirichlet priors to obtain preferences in word distribution characteristics however requires complicated priors that are capable of producing topic distributions that are not *iid*. The new prior will now need to generate a set of topics, which will no longer be independent of each other, instead of the Dirichlet prior from which multiple topics can be drawn in an iid manner.

When such priors are employed, they are no longer conjugate with the multinomial topic distributions necessitating sampling using computationally expensive methods like Metropolis-Hastings. The regularization technique described achieves a similar effect while requiring minimal additions to the existing Gibbs sampling inference procedure.

4 Injecting Labeled Features and Documents

In this section, we study how stronger prior knowledge in the form of labeled features and labeled documents can be incorporated into mixed-membership models by modifying the Gibbs sampling inference procedure. *Topic tables* are a commonly used method to display latent topics that are uncovered using models such as LDA. These tables depict topics using the top words of multinomials recovered after inference. Here, we use labeled features to indicate the topic a feature belongs to as a way to influence the formation of the topic tables. This is done by giving the inference procedure hints about the latent topic tables that we expect to see for the labeled features. Document labels, similarly bring the model closer to semi-supervised learning where a subset of the training data has known labels by providing apriori information about the latent topic assignment during inference.

Firstly, we look at a method to use labeled features by modifying the Gibbs sampler. As a concrete example, let us return to the task of clustering entities drawn from web tables. We might have domain knowledge that certain entities do not have multiple senses and should be assigned to a single pre-known latent cluster. An example is *Google* which in the context of our task is known to always be generated by the company topic. In general, we have pre-known latent cluster assignments for a small set of features which are strongly topic-indicative.

Let $\mathcal{L}$ be a set of pairs $\langle w, k_w \rangle$ where w is a feature i.e. $w \in V_t, t \in 1 \ldots T$ and $k_w \in 1, \ldots, K$. Each such pair indicates that the latent topic that generates an instance of w in the corpus is almost certainly k_w. Note that we do not have information about the nature of topic k_w at this stage before inference. We simply use the topic ids in $\mathcal{L}$ to separate and funnel features of different known clusters to different topics. During Gibbs sampling, when the topic indicator for a word is inferred, the procedure is modified to include a check to see if the word in question is present in $\mathcal{L}$. If yes, then instead of sampling a topic indicator for the word, the latent topic indicator is set to k_w with a probability of γ_f, where γ_f is a constant close to 1.

In terms of the generative story underlying LDA derived models, using labeled features implies that the topic multinomials $\beta_{t,k}$ are no longer drawn from the same symmetric Dirichlet priors parameterized by γ. Instead, the method implies that we use different asymmetric Dirichlet priors for each topic. For instance if $w \in V_t$ has a label k_w, then the prior for topic k_w is an asymmetric Dirichlet with parameters γ for all words other than w and a larger value γ^* for the word w. For all the other topics, the asymmetric Dirichlet has a lower value $\gamma^{'}$ for w to enforce our prior belief that w is more likely to be generated by topic k_w than any other topic.

Next, we examine how labeled data in the form of a-priori information about entity cluster membership that can be integrated into the inference procedure. While the motivation in using a LDA-derived approach for the entity clustering task lies in its ability to model mixed-membership, in the task of clustering entities, there are many entities that belong to only one cluster. In such a context, it would be useful to allow the inference procedure to use known cluster assignments for a small number of documents to influence the latent cluster formation. For instance, in the entity clustering task using the Toy Apple dataset, we might wish to use domain knowledge to say that "persimmon" belongs exclusively to the "fruit" cluster.

Let d be a document that is known to belong to cluster c_d. During inference using Gibbs sampling, for all words in the document, the cluster c_d is assigned with probability γ_d (≈ 1.0), and the usual Gibbs sampling procedure is used to determine the latent topic assignment with probability $1 - \gamma_d$.

Similar to the generative story underlying labeled features, the use of labeled documents implies a generative process where labeled documents' topic distributions i.e θ are drawn from asymmetric Dirichlet priors with higher parameter values for their topic labels instead of the symmetric Dirichlet priors that are usually used.

5 Experimental Results

First, we study the effect of biasing the model to better exploit topic-indicative features. Figure 2 shows the co-occurring entities perplexity of the biased Link-LDA model for the different datasets for different values of the variance parameter in the bias term. The reported values are averaged over 10 trials. For each trial, the Gibbs sampler ran for 100 iterations. The number of topics is set to 40 based on visual inspection of the clusters that were formed. The effect of regularization described below is however similar, when the number of topics is changed. It can be seen that the best perplexity is seen across all datasets when the variance is set to 0.2. We use this variance when using feature regularization (i.e. biasing) for the rest of the paper. When biasing is used, it is applied to the column id and entity-pair features: a column in a table is unlikely to contain entities from multiple clusters and is therefore strongly indicative of the topic; similarly, while an entity can belong to multiple topics, an entity-pair such as "apple:peach" is strongly indicative of a single topic.

Table 2 shows the difference in performance between the biased and baseline unbiased models as measured by NMI between predicted cluster distributions and known

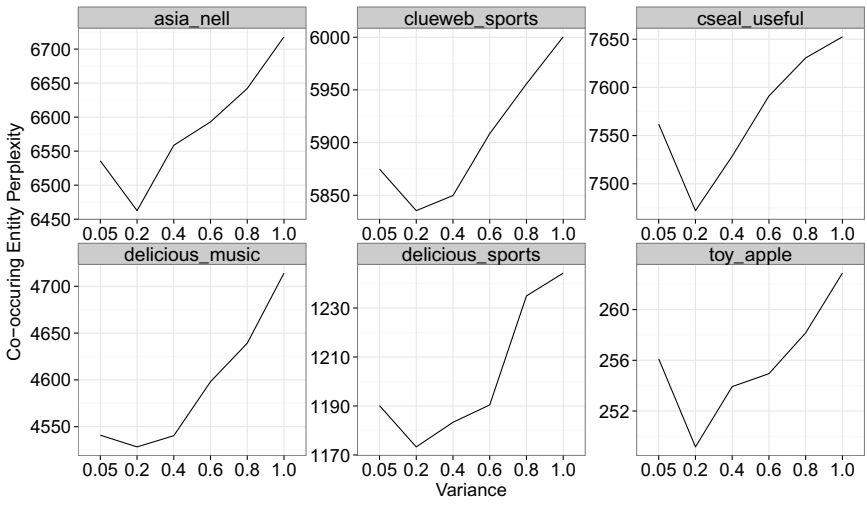

Fig. 2. Studying perplexity with feature regularization

Table 2. Feature regularization: Effect on NMI

Dataset	Regularization		Change
	No	Yes	
Asia-NELL	0.586	0.637	+8.70%
Clueweb-Sports	0.567	0.624	+10.05%
CSEAL-Useful	0.533	0.588	+10.31%
Delicious-Music	0.548	0.621	+13.32%
Delicious-Sports	0.609	0.615	+0.98%
Toy Apple	0.771	0.781	+1.29%

true cluster labels of labeled documents. For all the datasets, the biased models show a significant improvement over the unbiased variant. We note here that we cannot directly compare the entity clustering results from these experiments to the results from prior work in HTML table based entity clustering by Dalvi et al. [9] because the approach in that work clusters triplets of entities extracted from tables rather than individual entities. The biasing technique presented here is a general one and can be applied to any task that mixed-membership models are used for, whereas the WebSets approach specifically addresses the entity clustering task. For rough comparison however, the NMI value of clustering entities from the Delicious-Sports dataset is reported at 0.64 using the WebSets[9] approach whereas Table 2 indicates that the regularized model returns a NMI of 0.615 for the same dataset. It is worth re-emphasizing again that the results are not directly comparable.

Next, we study the effects of feature and document labeling in Figures 3 and 4. Feature and document labels are provided to the model for a subset of co-occurring entity features and entities. Labels for entities were obtained using Amazon's Mechanical Turk and were used to label entity documents and also co-occurring entity features. Although entities in general may have multiple senses, we only obtained labels for entities that have a single dominant sense. Table 3 shows the number of labeled features and documents for each dataset. In these figures, models are trained with increasing amount of supervision in the form of feature and document labels and the NMI between the true cluster labels of labeled documents and their inferred topic distributions for different model variants are plotted. It can be seen that as expected, increasing the amount of labeled data provided to the model results in higher NMI values for all model variants.

In figure 3, the red dashed line shows the performance of a mixture of multinomials (MoM) model[1] which allows each entity to belong to exactly one cluster. It can be seen that disallowing mixed-membership results in lower performance as compared to even the plain vanilla LDA model. The plot also indicates that the adding feature regularization (Link-LDA+FR) i.e. biased features consistently shows higher NMI values than the unbiased Link-LDA model and that adding all available document labels (Link-LDA+FR+DL) in addition to the different amounts of feature labels to the biased Link-LDA model yields the best NMI. It is interesting to note that adding feature

[1] While EM can be used for inference in the MoM model, we use Gibbs sampling for these experiments.

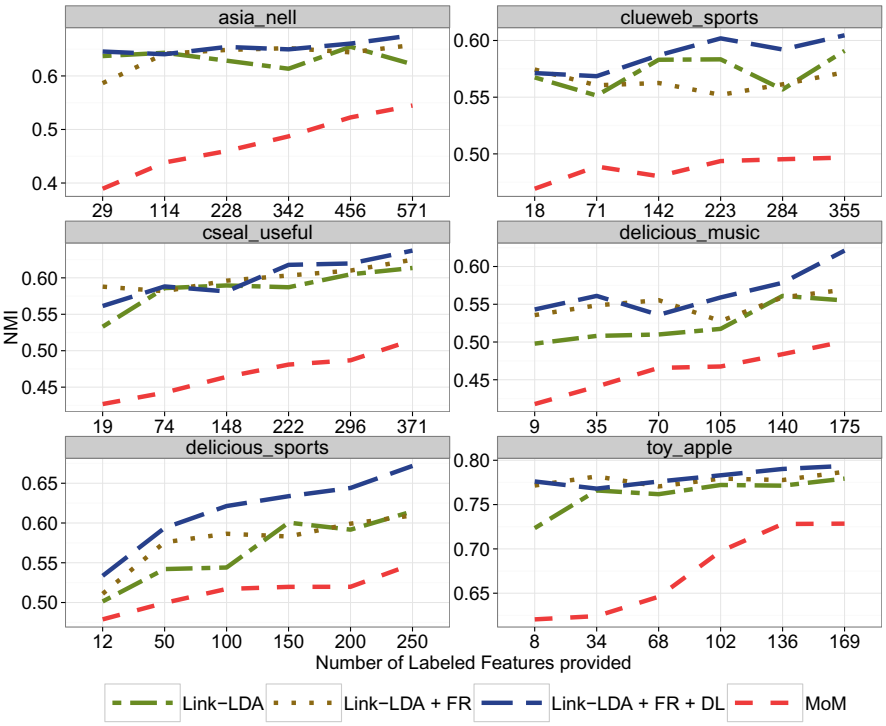

Fig. 3. Effect of injecting Feature Labels

labeling to the mixture of multinomials model, i.e., the points on the MoM line towards the right of the plot, describes a setting that is similar to DUALIST [21].

The entropy of θ can be subject to the same kind of regularization as the word topic distribution used in feature regularization, enabling us to restrict the degree to which entities are allowed to exhibit mixed-membership. In figure 4, it can be seen that adding such document regularization (+ DR), shows better performance than the regular Link-LDA model. Adding feature biasing (+ FR) and all available feature labels (+ FL), along

Table 3. Feature and Document Label statistics

Dataset	Co-occurring entities vocabulary size	#Labeled features	#Labeled documents
Asia-NELL	18309	571	411
Clueweb-Sports	28891	355	302
CSEAL-Useful	24240	371	600
Delicious-Music	9748	175	254
Delicious-Sports	3183	249	206
Toy Apple	423	169	177

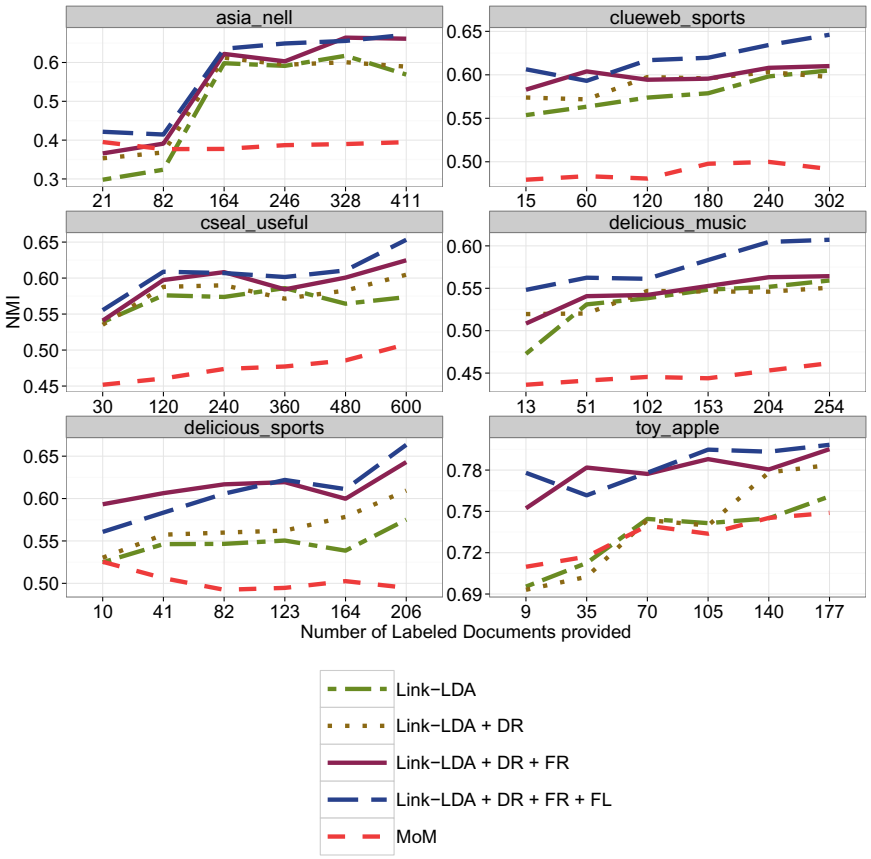

Fig. 4. Effect of adding Document Labels

with different degrees of document labels, shows progressively higher NMI across all datasets especially as the number of labeled documents provided is higher. The red dashed line in the plot representing the performance of the MoM model shows the performance of a single cluster membership model as we move from a fully unsupervised model to a semi-supervised model.

The above experiments show that introducing labeled documents and features consistently improves performance. While document labels have more impact, the labeling scheme used restricts us to only provide labels for entities with a single sense. We also see that for a fixed number of feature or document labels, adding feature regularization (i.e. biasing) and document regularization consistently improves the NMI scores.

In Table 4, we see illustrative examples of the advantage of the mixed-membership approach. For the ambiguous entities shown, the top two topics to which they are deemed to belong are shown using the top entries from the entity-pair multinomials. The results are from a biased Link-LDA model with no labeled features or documents. The topic titles in bold were added after inference by looking at the top entries for the

Table 4. Mixed-membership clustering results of ambiguous entities

Dataset: asia_nell, Entity: **franklin**
Names: (0.24) armstrong:brown, jennifer:jessica, chloe:gucci, brandon:joseph, benjamin:matthew, donald:edward, russell:stanley, benjamin:ethan, greg:gregg, angel:jose **Places:** (0.21) montana:nebraska, dakotas:north_carolina, rock_island:san_francisco, atlanta:long_island, delaware:montana, montana:new_york, central_california:san_clemente_island, clearwater:cocoa_beach, sutter:tehama, oklahoma_city:salt_lake_city
Dataset:toy apple, Entity: **apple**
Food: (0.61) peaches:pears, cocoa:coconut_oil, apricots:avocados, sodium_carbonate:sodium_chloride, lactic_acid:lauric_acid, sugar_alcohols:sugars, coconut_oil:coffee, caffeine:calcium_carbonate, xanthan_gum:yeast, sodium_citrate:sodium_hydroxide, pears:pineapple **Companies:** (0.16) nec:palmone, blackberry:google, sony:tomtom, asus:palm, philips:samsung, dell:ericsson, sagem:sharp, orange:philips, asus:google, sagem:samsung, asus:bosch
Dataset:delicious sports, Entity: **giants**
NFL teams: (0.26) chiefs:redskins, browns:raiders, cardinals:redskins, rams:saints, cowboys:redskins, cowboys:jaguars, bengals:eagles, bengals:patriots, falcons:patriots, saints:falcons, eagles:panthers **MLB teams:** (0.21) arizona_diamondbacks:cincinnati_reds, pittsburgh_pirates:texas_rangers, cleveland_indians:minnesota_twins, milwaukee_brewers:san_diego_padres, boston_red_sox:los_angeles_dodgers, cincinnati_reds:new_york_yankees, minnesota_twins:pittsburgh_pirates, florida_marlins:houston_astros, chicago_cubs:los_angeles_dodgers, baltimore_orioles:montreal_expos, houston_astros:philadelphia_phillies

topic. The value in parentheses show the degree of membership that the entity has for the topic. It can be seen that the mixed-membership latent variable model approach is able to detect the multiple senses of ambiguous entities. The first entity in the table **Franklin** is ambiguous because it has multiple senses — as a common first or last name and as a name of a city in the state of Nebraska in the US, among others. The second example **apple** as discussed earlier could either refer to the fruit or the company. The top two topics returned for this entity denotes exactly these two concepts. The third example **giants** is from the sports domain and could refer to either the New York Giants who play in the National Football League (American Football) or the San Francisco Giants who play in Major League Baseball (MLB). The top two topics indicate these two concepts.

6 Related Work

Ganchev et al. [11] proposed Posterior Regularization (PR), a method to incorporate indirect supervision via constraints on posterior distributions of probabilistic models with latent variables. They demonstrate the use of the technique in models for several tasks such as POS induction, word alignment, etc. While the approach proposed in this paper is similar in spirit to PR in that both approaches provide a method for preferences

for the posteriors of latent variables to be specified, there are significant differences. The PR framework is used in applications where exact inference is possible and the authors present a modified EM procedure to learn parameters for the model and incorporate constraints in an interleaved manner. In the approach introduced in this paper to bias the model, we focus on incorporating constraints on latent role distributions in models where exact inference is intractable by incorporating the constraints into the model instead of imposing them in a separate distinct step during inference.

Mann and McCallum [14] also proposed a general framework to introduce preferences in model expectations by adding terms called *generalized expectation (GE) criteria* to the objective function. Examples of such criteria were explored in the domain of log linear models. The approach in this paper is similar to the GE framework in that the regularization operates on entropies of distributions of inferred latent variables. The manner in which deviations from expectations are penalized, however differs from the criteria used by Mann and McCallum; the method introduced in this paper proposes that a desired value is drawn from a distribution parameterized by the inferred latent variables' values. The GE framework has not been applied to latent variable mixed-membership models as far as we know.

Newman et al. [16] presented a method to regularize topic models to produce coherent topics. In this approach, a pre-computed matrix of word-similarities from external data (Wikipedia) is used to construct a prior for the topic distributions. This regularization approach differs from the framework used in this paper in that it is aimed at producing topics that respect external word similarities. This is in contrast to our approach that is designed to control the latent structure properties without using external data.

Incorporating document labels into classifiers to obtain semi-supervised models is a well established technique in machine learning [17]. In the context of topic models, Labeled-LDA [20] uses tags attached to documents to limit the membership of the documents to specified topics. The labeled document injection technique discussed in this paper is closely related to Labeled-LDA. Supervised LDA [5] is a related model where supervision in the form of categorical or real-valued attributes of documents is provided. These attributes are derived from the topic distributions using regression models, which differs from the approach in this paper where the document labels directly indicate topic membership. Mimno et al. [15] propose a model where the Dirichlet prior for document topic proportion distribution is replaced with a log-linear prior that permits the distribution to be directly influenced by metadata. This work can be interpreted as a method to use metadata to tailor the latent structure formation. Settles [21] used labeled features for multinomial Naive Bayes classifiers. A similar approach was used by Attenberg et al. [3] in the context of active learning.

Steyvers et al. [22] present a related approach where they "pre-construct" some topics based on concepts obtained from Cambridge Advance Learner's Dictionary (CALD). This approach is similar to the labeled features idea presented in this paper. A concept topic as defined by this approach can be seen as a set of labeled words with the same topic indicator.

Entity clustering from semi-structured data has been addressed previously [18,23,9]. These approaches however do not address the issue of mixed-membership.

7 Conclusion

We presented a novel technique to bias latent variable mixed-membership models to exploit topic-indicative features and used the biased model for the task of clustering semi-structured data in the form of entities extracted from HTML tables. Our experiments show that the biased models outperform the baseline models in the cluster recovery task as measured by NMI. We then presented a method to allow for stronger supervision in the form of feature and document labels to move further along the spectrum toward semi-supervised learning from totally unsupervised learning. Results indicate that the stronger forms of supervision result in better cluster recovery. To summarize, we presented a framework in which mixed-membership models can be successfully used in a semi-supervised fashion to incorporate inexpensive weak prior domain knowledge.

References

1. Andrzejewski, D., Buttler, D.: Latent topic feedback for information retrieval. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2011, pp. 600–608. ACM, New York (2011), http://doi.acm.org/10.1145/2020408.2020503
2. Arora, R., Ravindran, B.: Latent dirichlet allocation and singular value decomposition based multi-document summarization. In: ICDM, pp. 713–718. IEEE Computer Society (2008)
3. Attenberg, J., Melville, P., Provost, F.: A unified approach to active dual supervision for labeling features and examples. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010, Part I. LNCS (LNAI), vol. 6321, pp. 40–55. Springer, Heidelberg (2010)
4. Balasubramanyan, R., Cohen, W.W.: Regularization of latent variable models to obtain sparsity. In: SDM (2013)
5. Blei, D., McAuliffe, J.: Supervised topic models. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems, vol. 20, pp. 121–128. MIT Press, Cambridge (2008)
6. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003), http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf
7. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr., E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: Proceedings of the Twenty-Fourth Conference on Artificial Intelligence, AAAI 2010 (2010)
8. Carlson, A., Betteridge, J., Wang, R.C., Jr. Hruschka, E.R., Mitchell, T.M.: Coupled semi-supervised learning for information extraction. In: Proceedings of the third ACM International Conference on Web Search and Data Mining, WSDM 2010, pp. 101–110. ACM, New York (2010), http://doi.acm.org/10.1145/1718487.1718501
9. Dalvi, B.B., Cohen, W.W., Callan, J.: Websets: extracting sets of entities from the web using unsupervised information extraction. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM 2012, pp. 243–252. ACM, New York (2012), http://doi.acm.org/10.1145/2124295.2124327
10. Erosheva, E.A., Fienberg, S., Lafferty, J.: Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences of the United States of America 101(suppl. 1), 5220 (2004)
11. Ganchev, K., Graça, J.A., Gillenwater, J., Taskar, B.: Posterior regularization for structured latent variable models. J. Mach. Learn. Res. 11, 2001–2049 (2010), http://dl.acm.org/citation.cfm?id=1756006.1859918

12. Griffiths, T.L., Steyvers, M., Blei, D.M., Tenenbaum, J.B.: Integrating topics and syntax. In: Advances in Neural Information Processing Systems, vol. 17, pp. 537–544. MIT Press (2005)
13. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistics, COLING 1992, vol. 2, pp. 539–545. Association for Computational Linguistics, Stroudsburg (1992), http://dx.doi.org/10.3115/992133.992154
14. Mann, G.S., Mccallum, A.: Generalized Expectation Criteria for Semi-Supervised Learning with Weakly Labeled Data. Journal of Machine Learning Research 11, 955–984 (2010), http://dl.acm.org/citation.cfm?id=1756038
15. Mimno, D.M., McCallum, A.: Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In: McAllester, D.A., Myllymäki, P. (eds.) UAI, pp. 411–418. AUAI Press (2008)
16. Newman, D., Bonilla, E.V., Buntine, W.L.: Improving topic coherence with regularized topic models. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F.C.N., Weinberger, K.Q. (eds.) NIPS, pp. 496–504 (2011)
17. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using em. Mach. Learn. 39(2-3), 103–134 (2000), http://dx.doi.org/10.1023/A:1007692713085
18. Paca, M., Van Durme, B.: Weakly-Supervised Acquisition of Open-Domain Classes and Class Attributes from Web Documents and Query Logs, pp. 19–27. Association for Computational Linguistics, Columbus (2008), http://www.aclweb.org/anthology/P/P08/P08-1003
19. Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., Welling, M.: Fast collapsed gibbs sampling for latent dirichlet allocation. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 569–577 (2008)
20. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 248–256. Association for Computational Linguistics, Singapore (2009)
21. Settles, B.: Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1467–1478. Association for Computational Linguistics, Edinburgh (2011), http://www.aclweb.org/anthology/D11-1136
22. Steyvers, M., Smyth, P., Chemuduganta, C.: Combining Background Knowledge and Learned Topics. Topics in Cognitive Science 3(1), 18–47 (2011), http://doi.wiley.com/10.1111/j.1756-8765.2010.01097.x
23. Talukdar, P.P., Reisinger, J., Pasca, M., Ravichandran, D., Bhagat, R., Pereira, F.: Weakly-Supervised Acquisition of Labeled Class Instances using Graph Random Walks, pp. 582–590. Association for Computational Linguistics, Honolulu (2008), http://www.aclweb.org/anthology/D08-1061
24. Wang, R.C., Cohen, W.W.: Automatic set instance extraction using the web. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and The 4th International Joint Conference on Natural Language Processing of the AFNLP, ACL 2009, vol. 1, pp. 441–449. Association for Computational Linguistics, Stroudsburg (2009), http://dl.acm.org/citation.cfm?id=1687878.1687941
25. Wetzker, R., Zimmermann, C., Bauckhage, C.: Analyzing social bookmarking systems: A del.icio.us cookbook. In: Mining Social Data (MSoDa) Workshop Proceedings, ECAI 2008, pp. 26–30 (July 2008), http://robertwetzker.com/wp-content/uploads/2008/06/wetzker_delicious_ecai2008_final.pdf