Constrained Relational Topic Models

Silvia Terragni[1], Elisabetta Fersini, Enza Messina

DISCo, University of Milano-Bicocca, Viale Sarca, 336 – 20126 Milano, Italy

Abstract

Relational topic models (RTM) have been widely used to discover hidden topics in a collection of networked documents. In this paper, we introduce the class of Constrained Relational Topic Models (CRTM), a semi-supervised extension of RTM that, apart from modeling the structure of the document network, explicitly models some available domain knowledge. We propose two instances of CRTM that incorporate prior knowledge in the form of document constraints. The models smooth the probability distribution of topics such that two constrained documents can either share the same topics or denote distinct themes. Experimental results on benchmark relational datasets show significant performances of CRTM on a semi-supervised document classification task.

Keywords:

Constrained Relational Topic Models, Semi-Supervised Model, Latent Dirichlet Allocation, Domain Knowledge.

1. Introduction

Probabilistic topic models are a promising class of generative probabilistic models that provide a simple way to analyze and summarize the main themes of large volumes of texts. A topic model describes a corpus of documents through a fixed set of topics, seen as distributions of words over a fixed vocabulary. Each document is represented as a "bag of words" and it is assumed as composed of

[1]Corresponding author.

Email addresses: s.terragni4@campus.unimib.it (S. Terragni), fersiniel@disco.unimib.it (E. Fersini), messina@disco.unimib.it (E. Messina)

a mixture of different topics, where a topic drawn from this mixture is assigned to each word of the document.

Classical topic models consider texts as the unique source of information and are generally based on the assumption that texts are independent and identically distributed (i.i.d.), do not taking into account that documents and its constituents (e.g. unigrams or n-grams) can actually show an underlying relational structure. In realistic cases, documents are often related to each other: scientific papers can be related through citations, web pages can present hyperlinks between each other, and users in social networks can be friends. In these networked environments, connected documents likely discuss the same topics. Relational information can therefore be exploited for uncovering and understanding the underlying latent structure of a corpus of documents.

Relational Topic Model (RTM) [12] is a hierarchical model that explicitly ties links between documents and their contents. As a topic model, it produces a low dimensional topical representation of a document that can be used to address different tasks, such as information retrieval, document exploration, and clustering. Recently, many topic models that include relational information between documents have been proposed [50, 47, 13, 44, 45, 23], some of them being direct extensions of RTM [13, 44, 45]. Most of them include only one type of relational information, i.e. the links between documents, in addition to the text, disregarding that documents can also provide some other prior knowledge: for example, a domain expert may know the label associated with a document or that two documents belong either to the same or a different class. The introduction of prior knowledge can further strengthen the probabilistic process of generating topics and words in order to improve the model's performance. The main contributions of this paper are the following:

- we propose a class of relational topic models, named Constrained Relational Topic Model (CRTM), that is a semi-supervised extension of RTM that includes not only the information about the network of documents, but it jointly models the available prior knowledge about documents in

the form of constraints;

- inspired by constraint-based semi-supervised clustering [6], we propose two instances of CRTM that include the knowledge at the document level in the form of must-link and cannot-link constraints between documents. The models can smooth the distribution of topics so that two constrained documents can either share the same topics or denote distinct themes.

The paper is organized as follows. In section 2, a brief overview of the state of the art about relational topic models is presented. In section 3, a brief review of Relational Topic Model (RTM) is given. In section 4, the proposed Constrained Relational Topic Model is detailed. In section 5, the experimental settings are described, while in section 6 the experimental results are discussed. Finally, in section 7, conclusions and future work are reported.

2. Related Work

Latent Dirichlet Allocation (LDA) [10] is a generative probabilistic model that describes a corpus of documents through a set of topics K, seen as distributions of words over a fixed vocabulary W. Each document is assumed composed of a mixture of different topics that follow a Dirichlet distribution, where a topic drawn from this mixture is assigned to each word of the document. LDA and simple topic models can be extended by considering different types of relational information, originating the following relational topic models.

Word-Level Relational Topic Models. This type of relational topic model relaxes the independence assumption of words in a document or in a topic. Different types of underlying relationships between words can be considered, as word-order [21, 43, 22, 31, 17, 40], syntactic dependencies [20, 11], semantic or domain knowledge relationships [2, 3, 13, 46]. Hidden Markov Topic Model [22] assumes that words in each sentence are assigned to the same topic. Generalized Pòlya Urn Model [17] is a model based on the homonymous process which considers not only every single word of a document, but also collocations, to

produce more understandable topics. Syntactic Topic Model [11] is a combination of a topic model and syntax model, assuming that a topic assignment for a word depends on both a document-level observation and parse tree-level information. In [3], the authors present a topic model which incorporates domain knowledge by using First-Order Logic rules to represent relationships between words. General Knowledge based LDA [14] incorporates lexical semantic relationships of words (such as synonyms and antonyms) by adding a latent variable which denotes the relation for each word. Constrained LDA [46] can incorporate relationships among words through the use of a potential function which smooths the probability distribution of topics.

In the latest years, the increasing interest on word embeddings has led to the incorporation into topic models of the semantic relationships that these distributed representation of words are able to capture [37, 48, 16, 35, 28, 7]. For example, in [16] the proposed model directly generates word embeddings, instead of discrete word types, using a multivariate Gaussian distribution. In [35] the word-topic multinomial distribution is replaced by a Dirichlet multinomial distribution and a latent word feature distribution.

Document-Level Relational Topic Models. The main paradigms of relational models that consider the underlying network structure of a collection of documents are Relational Topic Models (RTM), Regularized Topic Models, Dirichlet Multinomial Regression (DMR), and Bayesian Deep Learning.

RTM [12] and its extensions are based on LDA and model each link as a binary variable, thus considering the existence (or absence) of a link between a couple of documents. Generalized RTM [13] can capture not only same-topic relationships between documents but all pairwise topic relationships. Sparse RTM [47] aims at inferring sparse topics for each document by using a non-probabilistic formulation of RTM. In [44] and [45] RTM is turned into a supervised topic model, where the link is the variable to predict.

Regularized topic models [24, 33] aim to augment the topic model objective function with a network regularization penalty that encourages topic mixtures

of related documents to be similar.

Dirichlet Multinomial Regression (DMR) [34] and its extensions [25, 39] are topic models that incorporate arbitrary features, considering links as per-document attributes.

In the latest years, a new paradigm has been studied, that aims to combine deep learning and probabilistic models in a unified framework: in [42] a Bayesian deep learning framework jointly models high-dimensional node attributes and link structures with layers of latent variables, and in [5] topics are inferred using a Stacked Variational AutoEncoder and the latent representations of a pair of documents are concatenated as the input of a multilayer perceptron that predicts a link.

Topic-Level Relational Topic Models. LDA assumes topics are independent of each other, however, in a realistic application, this assumption can be too simplistic and restrictive. Models that aim to express the interactions between topics usually adjust Dirichlet priors α or β, which generate the document-topic distribution and the word-topic distribution respectively.

In particular, Correlated Topic Model [9] allows pairwise correlations between topic, Latent Dirichlet-Tree Allocation [38] replaces the Dirichlet prior α with a Dirichlet-Tree distribution to express hierarchies of topics, and Pachinko Allocation Model [29] represents the relationships among topics as arbitrary directed acyclic graphs. Several models [36, 27, 1] extend the nested Chinese Restaurant Process (nCRP) [8] by introducing a tree structure prior constructed with multiple CRPs.

In recent years, Poisson Factor Analysis (PFA), a nonnegative matrix factorization model with Poisson link, has been extended to its deep counterpart, leading to models that infer topics through deep latent hierarchies [18, 26, 49, 15].

The class of topic models that we propose belongs to the family of Document-Level Relational Topic Models. Rather than focusing on the representation of

documents, we focus our investigation on the encoding of document relationships to incorporate domain knowledge. In particular, we propose a semi-supervised extension of RTM that includes not only the document words and the relational information about the network of documents, but it jointly models the available prior knowledge at the document level in the form of constraints.

3. Relational Topic Model

Since the class of topic models that we propose is built upon Relational Topic Model (RTM), in this section we briefly review the model.

Relational Topic Model extends Latent Dirichlet Allocation by modeling a link between a pair of documents using a link probability function that depends on the topic assignments $\mathbf{z}$ of the considered documents, thus assuming that documents with similar topic assignments are likely to be linked.

The link likelihood function can be defined in different ways; in this paper, we consider the sigmoid function, parameterized by coefficient η and intercept ν. The likelihood that a link y between two documents d and d' exists is then computed as:

$$\psi_\sigma(y = 1) = \sigma(\eta^T(\overline{\mathbf{z}}_d \circ \overline{\mathbf{z}}_{d\prime}) + \nu)$$

In particular, σ is the sigmoid function, the symbol $\circ$ denotes the Hadamard product (or element-wise product) and $\overline{\mathbf{z}}_d$ is a vector, such that $\overline{\mathbf{z}}_d = \frac{1}{N_d}\sum_{n=1}^{N_d} z_{nd}$, where N_d is the length of document d and z_{nd} denotes the n-th word in document d.

Being an extension of LDA, the joint probability distribution of RTM is composed by the joint distribution of LDA and the term related to the links

between documents:

$$p(\mathbf{w}, \mathbf{z}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\Phi} | \alpha, \beta, \eta, \nu) \tag{1a}$$

$$= P(\boldsymbol{\theta}|\alpha) P(\mathbf{w}|\mathbf{z}, \boldsymbol{\phi}) P(\mathbf{z}|\boldsymbol{\theta}) P(\boldsymbol{\phi}|\beta) \psi_\sigma(\mathbf{y}|\mathbf{z}, \mathbf{w}, \eta, \nu) \tag{1b}$$

$$= \prod_{d}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(w_{nd}|\phi_{z_{nd}}) p(z_{nd}|\theta_d) \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{\substack{d,d' \in D \\ d' \neq d}} \psi_\sigma(y_{dd'}|z_d, z_{d'}, \eta, \nu) \tag{1c}$$

where

- D denotes the set of documents
- N_d is the number of words of document d
- K denotes the fixed number of topics
- $\mathbf{w}$ denotes the set of words and w_{nd} denotes the n-th word in document d
- $\mathbf{z}$ represents the set of topic assignments and z_{nd} the topic assignment of the n-th word in document d
- $\mathbf{y}$ is the link variable denoting the existence or absence of a link between two documents and $y_{dd'}$ denotes the link between documents d and d'
- $\boldsymbol{\theta}$ represents the document-topic distribution and θ_d the distribution of topics for document d
- $\boldsymbol{\phi}$ denotes the word-topic distribution and $\phi_{z_{nd}}$ is the distribution related to the topic assignment of the n-th word in document d
- α and β are the Dirichlet hyper-parameters related to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ respectively.

4. Constrained Relational Topic Models

Most of Document-Level Relational Topic Models consider only the document network information, disregarding that other types of information deriving

from domain knowledge can be encoded as well. Building upon RTM, we introduce prior knowledge at the document level in the form of constraints through the definition of a set of potential functions, inspired by the Word-Level Relational Topic Model Constrained-LDA [46].

The prior knowledge is denoted by a set L and each knowledge $l \in L$ is introduced into the model by a potential function $f_l(z, d)$, which represents a real-valued score for the hidden topic assignment z in document d. The complete prior knowledge L defines a score $\xi(\mathbf{z}, L) = \prod_{z \in \mathbf{z}} \exp f_l(z, d)$ that smooths the current topic assignment $\mathbf{z}$. The joint probability distribution of this class of topic models, to which we will refer to as Constrained Relational Topic Models (CRTM), is defined as follows:

$$p(\mathbf{w}, \mathbf{z}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\phi} | \alpha, \beta, \eta, \nu, L) \tag{2a}$$

$$= P(\mathbf{w}|\mathbf{z}, \boldsymbol{\phi}) P(\boldsymbol{\phi}|\beta) P(\mathbf{z}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\alpha) \psi_\sigma(\mathbf{y}|\mathbf{z}, \mathbf{w}, \eta, \nu) \xi(\mathbf{z}, L) \tag{2b}$$

The potential function ξ and the link probability function ψ_σ can be factored out of the marginalized joint distribution, because they do not depend on the distributions ϕ and θ, obtaining the following marginalized joint probability distribution:

$$p(\mathbf{w}, \mathbf{z}, \mathbf{y} | \alpha, \beta, \eta, \nu, L) \tag{3a}$$

$$= \int \int p(\mathbf{w}|\mathbf{z}, \boldsymbol{\phi}) p(\boldsymbol{\phi}|\beta) p(\mathbf{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\alpha) \psi_\sigma(\mathbf{y}|\mathbf{z}, \mathbf{w}, \eta, \nu) \xi(\mathbf{z}, L) d\theta d\phi \tag{3b}$$

$$= \xi(\mathbf{z}, L) \psi_\sigma(\mathbf{y}|\mathbf{z}, \mathbf{w}, \eta, \nu) \int \int p(\mathbf{w}|\mathbf{z}, \boldsymbol{\phi}) p(\boldsymbol{\phi}|\beta) p(\mathbf{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\alpha) d\theta d\phi \tag{3c}$$

The main goal of CRTM is to estimate the posterior distribution $P(\mathbf{z}|\mathbf{w}, \mathbf{y}) = P(\mathbf{w}, \mathbf{z}, \mathbf{y}) / \sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z}, \mathbf{y})$. Since the evaluation of the denominator is intractable, an approximate inference method is mandatory.

In our investigation, we use a collapsed Gibbs sampler that leads to the

following estimation:

$$P(z_{nd}|w_{nd}, \mathbf{z}^{\neg nd}, \mathbf{y}, \alpha, \beta, \eta, \nu, L) \tag{4a}$$

$$= \frac{P(\mathbf{w}, z_{nd}, \mathbf{z}^{\neg nd}, \mathbf{y}|\alpha, \beta, \eta, \nu, L)}{P(\mathbf{w}, \mathbf{z}^{\neg nd}, \mathbf{y}|\alpha, \beta, \eta, \nu, L)} \tag{4b}$$

$$= \frac{P(\mathbf{w}, z_{nd}, \mathbf{z}^{\neg nd}|\alpha, \beta)}{P(\mathbf{w}, \mathbf{z}^{\neg nd}|\alpha, \beta)} \prod_{\substack{d' \neq d \\ y_{dd'}=1}} \frac{\psi_\sigma(y_{dd'} = 1|z_{nd}, \mathbf{z}_d^{\neg nd}, \mathbf{z}_{d'}, \eta, \nu)}{\psi_\sigma(y_{dd'} = 1|\mathbf{z}_d^{\neg nd}, \mathbf{z}_{d'}, \eta, \nu)} \tag{4c}$$

$$\cdot \prod_{\substack{d' \neq d \\ y_{dd'}=0}} \frac{\psi_\sigma(y_{dd'} = 0|z_{nd}, \mathbf{z}_d^{\neg nd}, \mathbf{z}_{d'}, \eta, \nu)}{\psi_\sigma(y_{dd'} = 0|z_d^{\neg nd}, \mathbf{z}_{d'}, \eta, \nu)} \cdot \frac{\xi(\mathbf{z}^{\neg nd}, \mathbf{z}_{nd}, L)}{\xi(\mathbf{z}^{\neg nd}, L)} \tag{4d}$$

$$\propto \left(N_{dz_{nd}}^{\neg nd} + \alpha\right) \frac{N_{z_{nd}w}^{\neg nd} + \beta}{N_{z_{nd}\cdot} + W\beta} \tag{4e}$$

$$\cdot \prod_{\substack{d' \neq d \\ y_{dd'}=1}} \sigma\left(\frac{\eta_{z_{nd}}}{N_{d\cdot}} \cdot \frac{N_{d'z_{nd}}}{N_{d'\cdot}} + \sum_{k=1}^{K} \eta_k \frac{N_{dk}^{\neg nd}}{N_{d\cdot}} \frac{N_{d'k}}{N_{d'\cdot}} + \nu\right) \tag{4f}$$

$$\cdot \prod_{\substack{d' \neq d \\ y_{dd'}=0}} 1 - \sigma\left(\frac{\eta_{z_{nd}}}{N_{d\cdot}} \cdot \frac{N_{d'z_{nd}}}{N_{d'\cdot}} + \sum_{k=1}^{K} \eta_k \frac{N_{dk}^{\neg nd}}{N_{d\cdot}} \frac{N_{d'k}}{N_{d'\cdot}} + \nu\right) \cdot \exp\left(f_l(z_{nd}, d)\right) \tag{4g}$$

where

- the superscript $\neg nd$ indicates leaving the nth token of the dth document out of the calculation
- W represents the number of unique words in the vocabulary
- N_{dz} denotes the number of words associated with the topic z in document d
- $N_{z\cdot}$ denotes the number of words associated with the topic z in the corpus
- N_{zw} denotes the number of occurrences of the word w associated with topic z.

Equation (4e) corresponds to the Gibbs sampling of the standard LDA [19], equation (4f) represents the sigmoid link likelihood function of RTM when a link exists, and equation (4g) denotes the link function of RTM when a link is absent,

plus the incorporation of prior knowledge by means of the potential function. Let us notice that equations (4f) and (4g) related to RTM deal with directed graphs, however it can be easily adapted to deal with undirected networks, similarly to all the models that extend RTM.

4.1. Document Constraint Potential Functions

The introduction of document constraints instead of document labels can be more realistic in some cases. For instance, as observed in [6], labels may be unknown, but a user may know whether two documents belong or do not belong to the same class. This formulation is also more general, as document constraints imply labels, but the vice versa does not hold. In this paper, we propose two potential functions, inspired by must-link and cannot-link constraints described in [2], that allow us incorporating document constraints in RTM.

We define two knowledge sets for each document d: a *must-constraint* set L_d^m, containing documents that must share the same topics of d, and a *cannot-constraint* set L_d^c, including documents that cannot share the same themes of d. For example, a must-constraint set for the book titled *Emma* and written by Jane Austen could be

$$L_{Emma}^m = \{\textit{Sense and Sensibility}, \textit{Pride and Prejudice}\}$$

which contains a set of books written by the same author. Analogously, a cannot-constraint set could be

$$L_{Emma}^c = \{\textit{Moby-Dick}\}$$

which denotes a book that has not been written by Jane Austen.

In the following, we will detail two potential functions, which once instantiated in CRTM will lead to Unnormalized and Normalized CRTM.

4.1.1. Unnormalized Constrained Relational Topic Model (CRTM-U)

We can encode document relationships modeling the relationship that exists between the words of two constrained documents. In particular, we assume

that if two documents are must-constrained (i.e, they must share the same set of topic assignments) then the words in the documents must have similar topic distributions, i.e. $p(z_d|w,d) \approx p(z_{d'}|w',d')$, where w are the words of document d, and w' are the words of document d'. In order words, we model the idea that the more the words of the documents belonging to the set L_d^m are assigned to topic t, the higher the value of the potential function $f_l(z=t,d)$ is. Analogously, a cannot-constraint between two documents indicates that their words should not share the same set of topics. Therefore, if many words of two cannot-constrained documents are assigned to the same topic, then the value of the potential function will be low.

In order to model the previous ideas, we define the following potential function, named *unnormalized potential function*, as it takes into account the absolute value of the document-topic counts. It is defined as follows:

$$f_l(z,d) = \sum_{\substack{d' \in D \\ d' \in L_d^m}} \log \max\left(\lambda, N_{d'z}\right) + \sum_{\substack{d' \in D \\ d' \in L_d^c}} \log \frac{1}{\max\left(\lambda, N_{d'z}\right)} \tag{5}$$

where λ is the hyper-parameter which controls the strength of each $l \in L$. Larger values of λ imply that the constraint is active only for those topic assignments that have large counts. The value of λ must be set for each piece of knowledge according to the domain expert's confidence.

The conditional probability of topic z, including the defined document constraint potential function, can be estimated as:

$$P(z_{nd}|w,\mathbf{z}^{\neg nd},\mathbf{y},\alpha,\beta,\eta,\nu,L) \propto \left(N_{dz_{nd}}^{\neg nd} + \alpha\right) \frac{N_{z_{nd}w}^{\neg nd} + \beta}{N_{z_{nd}\cdot} + W\beta} \tag{6a}$$

$$\cdot \prod_{\substack{d' \neq d \\ y_{dd'}=1}} \sigma\left(\frac{\eta_{z_{nd}}}{N_{d\cdot}} \cdot \frac{N_{d'z_{nd}}}{N_{d'\cdot}} + \sum_{k=1}^{K} \eta_k \frac{N_{dk}^{\neg nd}}{N_{d\cdot}} \frac{N_{d'k}}{N_{d'\cdot}} + \nu\right) \tag{6b}$$

$$\cdot \prod_{\substack{d' \neq d \\ y_{dd'}=0}} 1 - \sigma\left(\frac{\eta_{z_{nd}}}{N_{d\cdot}} \cdot \frac{N_{d'z_{nd}}}{N_{d'\cdot}} + \sum_{k=1}^{K} \eta_k \frac{N_{dk}^{\neg nd}}{N_{d\cdot}} \frac{N_{d'k}}{N_{d'\cdot}} + \nu\right) \tag{6c}$$

$$\cdot \prod_{\substack{d' \in D \\ d' \in L_d^m}} \max\left(\lambda, N_{d'z_{nd}}\right) \prod_{\substack{d' \in D \\ d' \in L_d^c}} \frac{1}{\max\left(\lambda, N_{d'z_{nd}}\right)} \tag{6d}$$

The selection of the correct values for λ is not trivial, due to the different lengths of the documents that are involved in a constraint. For example, if we choose a value for λ that is too large, a document with a number of words less than λ will not affect the probability $p(z_{nd} = t \,|\, w_{nd}, \mathbf{z}^{\neg nd}, \mathbf{y}, L)$, even if all the words of the documents are assigned to topic t.

In order to smooth the effect of the hyper-parameter λ, a potential function that takes into account the length of the document is proposed in the next section.

4.1.2. Normalized Constrained Relational Topic Model (CRTM-N)

The following potential function considers the proportion of words in a document assigned to the same topic, rather than the absolute values of the document-topics counts. We define the potential function $f_l(z, d)$ as follows:

$$f_l(z,d) = \sum_{\substack{d' \in D \\ d' \in L_d^m}} \log\left(\frac{N_{d'z}}{N_{d'\cdot}} + 1\right) - \sum_{\substack{d' \in D \\ d' \in L_d^c}} \log\left(\frac{N_{d'z}}{N_{d'\cdot}} + 1\right) \tag{7}$$

The conditional probability of topic z estimated by CRTM, including the defined document constraint potential function, can be specified as follows:

$$P(z_{nd}|w, \mathbf{z}^{\neg nd}, \mathbf{y}, \alpha, \beta, \eta, \nu, L) \propto \left(N_{dz_{nd}}^{\neg nd} + \alpha\right) \frac{N_{z_{nd}w}^{\neg nd} + \beta}{N_{z_{nd}\cdot} + W\beta} \tag{8a}$$

$$\cdot \prod_{\substack{d' \neq d \\ y_{dd'}=1}} \sigma\left(\frac{\eta_{z_{nd}}}{N_{d\cdot}} \cdot \frac{N_{d'z_{nd}}}{N_{d'\cdot}} + \sum_{k=1}^{K} \eta_k \frac{N_{dk}^{\neg nd}}{N_{d\cdot}} \frac{N_{d'k}}{N_{d'\cdot}} + \nu\right) \tag{8b}$$

$$\cdot \prod_{\substack{d' \neq d \\ y_{dd'}=0}} 1 - \sigma\left(\frac{\eta_{z_{nd}}}{N_{d\cdot}} \cdot \frac{N_{d'z_{nd}}}{N_{d'\cdot}} + \sum_{k=1}^{K} \eta_k \frac{N_{dk}^{\neg nd}}{N_{d\cdot}} \frac{N_{d'k}}{N_{d'\cdot}} + \nu\right) \tag{8c}$$

$$\cdot \prod_{\substack{d' \in D \\ d' \in L_d^m}} \left(1 + \frac{N_{d'z_{nd}}}{N_{d'\cdot}}\right) \prod_{\substack{d' \in D \\ d' \in L_d^c}} \frac{1}{1 + \frac{N_{d'z_{nd}}}{N_{d'\cdot}}} \tag{8d}$$

In the following sections, an experimental investigation is presented to evaluate the capabilities of CRTM to discover hidden topics in different collections of networked documents.

5. Experimental Settings

In order to evaluate the performance of the proposed models, several experiments have been conducted, using document labels as prior knowledge and comparing their performance to different baseline models' performance.

5.1. Baseline Models

CRTM-N and CRTM-U have been validated comparing the results on benchmark datasets against the following models:

- LDA: Latent Dirichlet Allocation [10] using collapsed Gibbs sampling. Its joint probability distribution is as follows:

$$p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\Phi}|\alpha, \beta) = \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(w_{nd}|\Phi_{z_{nd}}) p(z_{nd}|\theta_d) \prod_{k=1}^{K} p(\Phi_k|\beta) \quad (9)$$

- RTM: standard RTM [12] that models only the links between documents through the binary variable $\mathbf{y}$, without incorporating any other kind of domain knowledge. The joint probability distribution corresponds to equation (1).

- Bi-RTM: RTM for bidimensional networks, where the first dimension is intended to represent the links of the document network, modeled by the binary variable $\mathbf{y}$, and the second dimension is designed to represent the must- and cannot-constraints between documents, modeled by an additional binary variable $\mathbf{c}$. Its joint probability distribution is the following:

$$\begin{aligned} & p(\mathbf{w}, \mathbf{z}, \mathbf{y}, \mathbf{c}, \boldsymbol{\theta}, \boldsymbol{\Phi}|\alpha, \beta, \eta, \nu, \eta', \nu') \\ & \quad = \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(w_{nd}|\Phi_{z_{nd}}) p(z_{nd}|\theta_d \prod_{k=1}^{K} p(\Phi_k|\beta) \\ & \quad \cdot \prod_{\substack{d,d' \in D \\ d' \neq d}} \psi_\sigma(y_{dd'}|z_d, z_{d'}, \eta, \nu) \cdot \prod_{\substack{d,d' \in D \\ d' \neq d}} \psi_{\sigma'}(c_{dd'}|z_d, z_{d'}, \eta', \nu') \end{aligned} \quad (10)$$

where η' and ν' are respectively the coefficient and the intercept for the sigmoid function $\psi_{\sigma'}$ that models the likelihood that a constraint $c_{dd'}$ between two documents d and d' exists.

For the sake of completeness, we also compared the proposed semi-supervised CRTM with a fully supervised model. In particular, we considered the following model:

- LLDA: Labeled LDA associates each label with one topic in direct one-to-one correspondence. In particular, we model this correspondence with the function $g : \Gamma \mapsto K$ that maps a label to its corresponding topic, where Γ is the set of labels. LLDA is built upon LDA and it is modeled by using the following potential function:

$$f_l(z, d) = \begin{cases} 1 & \text{if } z = g(l_d) \\ -\infty & \text{otherwise} \end{cases} \tag{11}$$

  where $l_d \in \Gamma$ specifies the document label.

Let us notice that in some realistic cases, we only know that two documents belong or do not belong to the same class, rather than knowing that to which class a document belongs.

5.2. Benchmark Datasets

Since CRTM deals with networks of documents and prior knowledge, the chosen datasets for the validation phase must have two main features: an underlying document-relational structure (e.g. citation links) and some domain knowledge available (e.g. document labels) to derive the semi-supervised constraints. Table 1 contains some statistics about the selected benchmarks. Cora [32] and M10 [30] are two datasets composed of 2708 and 4427 scientific publications respectively, whose links are represented by citations. WebKB[2] is a dataset composed of 877 universities web pages whose relationships are hyperlinks from a web page to another.

The three benchmarks have been preprocessed: words are stemmed, stopwords and the least and most frequent words are removed. Only the documents

[2] `http://www.cs.cmu.edu/~WebKB/ILP-data.html`

Dataset	#docs	#links	D ensity	Type of link	#classes	#unique words
Cora	2708	5430	$2.87 \cdot 10^{-4}$	citation	7	1752
M10	4427	5627	$7.41 \cdot 10^{-4}$	citation	9	1592
WebKB	877	1131	$14.72 \cdot 10^{-4}$	hyperlink	5	1830

Table 1: Statistics of the benchmark datasets Cora, WebKB, and M10.

that link another document or are linked by a document at least once are considered. Prior knowledge has been introduced in terms of constraints using a percentage of the possible constraints between documents. In particular, if two documents d and d' randomly chosen share the same class label, we expect that their words are assigned to similar topics, therefore a must-constraint is introduced (i.e. document d is added to the must-constraint set of d' and document d' is added to the must-constraint set of d). Concerning LLDA, we first define a mapping between the set of topics and the set of labels. Two documents d and d' are randomly drawn and the labels l_d and $l_{d'}$ are incorporated as knowledge, according to equation (11).

5.3. Performance Measures

Each dataset has been divided into a training set and a test set. The models are evaluated on the test set by measuring their performance on a document classification task. The K-dimensional representation of each document output by the considered topic model, i.e. the document-topic distribution θ, is used to train a linear Support Vector Machine (SVM) classifier that predicts the document classes. For the experimental evaluation, we considered both micro-F1 and macro-F1 measures.

Given a multi-class problem, F-measure, or F1 score, for a given class i is the weighted average of the precision and recall, and it reaches its best value at 1 and its worst score at 0. F-measure of class i is then defined as:

$$\text{f-measure}(i) = \frac{2 \cdot \text{Recall}(i) \cdot \text{Precision}(i)}{\text{Recall}(i) + \text{Precision}(i)}$$

Macro-F1. The average of the F1 score for each class is usually referred to as Macro-F1 or Macro-average F1 score. It is then defined as follows:

$$\text{Macro-F1} = \frac{1}{|Classes|} \sum_{i \in Classes} \text{f-measure}(i)$$

where $Classes$ denotes the set of the classes.

Micro-F1. The weighted average of the F1 score for each class (where the weight corresponds to the size of the classes) is called Micro-F1 or Micro-average F1 score, and it is then defined as:

$$\text{micro-average f-measure} = \frac{1}{D} \sum_{i \in Classes} |i| \cdot \text{f-measure}(i)$$

where D is the number of instances in the test set and $|i|$ the cardinality of class i.

5.4. Parameter settings

Each experiment, with a given set of parameters, has been repeated for 100 times and the performance measures have been averaged by the number of the samples, thus obtaining an average micro-F1 and macro-F1 measure. The hyper-parameters α and β have been set equal to $50/K$ and 0.1 respectively (as reported in [19]), for all the considered models. The selected value of λ for CRTM-U is 1. Each model has been trained for 1,500 Gibbs iterations.

The models have been validated by setting the number of topics equal to the number of classes of the dataset and by varying the quantity of prior knowledge, i.e. the number of possible constraints, during the training phase and, in a second stage, during the testing phase.

The maximum quantity of prior knowledge in terms of constraints is $\frac{D(D-1)}{2}$ (where D is the number of documents), which represents the maximum number of possible pairs among all the documents of the dataset. The quantity of knowledge introduced into the models is expressed as a percentage, preferring low values to maintain the typical semi-supervised scenario. Thus, given a percentage p, the number of constraints introduced into the model will be $p \cdot \frac{D(D-1)}{2}$,

rounded down to the nearest integer. When the percentage of knowledge is equal to 0%, then CRTM and LLDA correspond to RTM and LDA respectively.

We used Support Vector Machines (SVM) to predict the ground truth labels from the document-topic distribution of the documents. In particular, we used the LibSVM implementation[3] for inducing the linear SVM classifier.

The code of the proposed models is available at `https://github.com/MIND-Lab/Constrained-RTM`.

6. Experimental Results

In the following, we consider the performance of each model with an increasing percentage of prior knowledge introduced only in the training phase. In particular, we considered an experimental setting with zero knowledge (0.0%), which corresponds to models that do not encode any constraint (i.e. LDA and RTM) and represented in the plots using the lines. The other models, i.e. BiRTM, CRTM-U, CRTM-N, and LLDA, are reported with different percentages of knowledge, and they are represented by the bar plots.

Let us notice that in these experiments zero knowledge is incorporated in the testing phase, as it often happens in realistic cases.

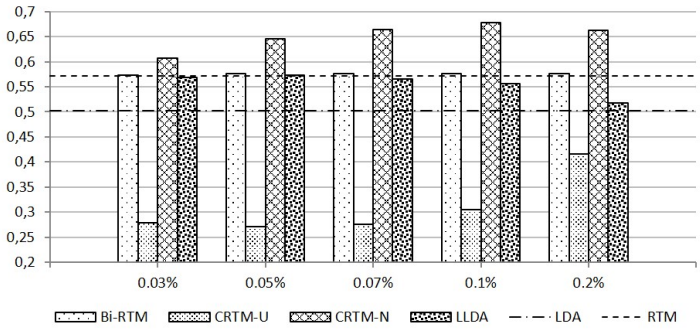

Figure 1: Micro-F1 performance of the compared models on Cora.

[3]LibSVM library: `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`

Figure 1 shows the performance of the models measured using the micro-F1 on the dataset Cora, where the number of constraints that are randomly selected ranges from 0.03% and 0.2% of the number of possible constraints. CRTM-N outperforms the other models, increasing its performance as more quantity of prior knowledge is introduced, and it seems to decrease its performance for larger quantities of constraints. We can also notice that while the performance of Bi-RTM is invariant with respect to the quantity of domain knowledge, LLDA gets at first an improvement with a small contribution of knowledge, then its performance decreases for larger values. The performance of CRTM-U is worse than the baselines LDA and RTM. The behavior of the model can be explained by the fact that documents in Cora are long (average length of a document is 68.9 words), thus the value of the potential function, which depends on the number of words associated with the current topic, will be very high allowing a small contribution to the rest of the sampling.

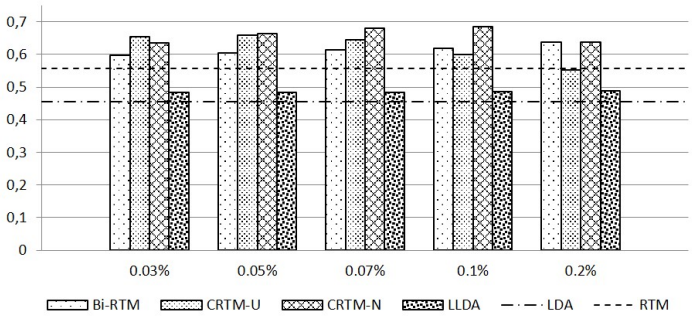

Figure 2: Micro-F1 performance of the compared models on M10.

In Figure 2, the results for dataset M10 are shown. CRTM-N has a similar behavior with respect to the previous experiments, while CRTM-U gets an improvement with a small insertion of constraints. This is due to the lengths of the documents of M10, which are short (the average length of documents in M10 is 6.3 words), thus making the introduction of the constraints more smoothed rather than in Cora. However, for larger quantities of knowledge, the average performance of CRTM-U gets worse. The introduction of the labels

allows LLDA to obtain a small improvement with respect of LDA, meaning that associating each word of a labeled document to the same topic does not improve the generalization capabilities of the model. Bi-RTM has a higher performance as the quantity of knowledge increases, although it requires many constraints and its best performance is still lower than CRTM-N.

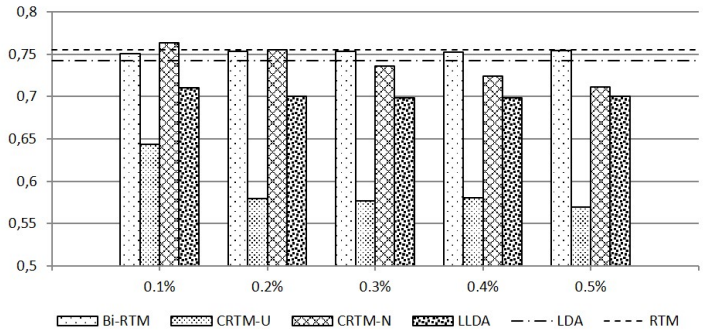

Figure 3: Micro-F1 performance of the compared models on WebKB.

Figure 3 shows the performance of the models for the dataset WebKB. CRTM-N still has the same behavior as the previous datasets, obtaining the best performance. Bi-RTM has a constant trend, while the other two models get worse performances with respect to LDA and RTM. The behavior of CRTM-U is similar to the one obtained in Cora. In fact, also WebKB is composed of long documents. On the other hand, LLDA has a lower performance with respect to Bi-RTM, CRTM-N, and LDA.

We report in the following the results of the considered models on the different datasets by introducing domain knowledge both in the training and testing set. In particular, each combination of values of percentage in training and testing has been considered. To provide a concise visualization of the performance of the models, the results have been averaged, and therefore Figure 4 reports the best average performance for each model.

The two CRTMs significantly outperform Bi-RTM and the baselines LDA and RTM (with a confidence of 95%). In particular, the two proposed models have similar performance on M10 and WebKB, while CRTM-N outperforms its

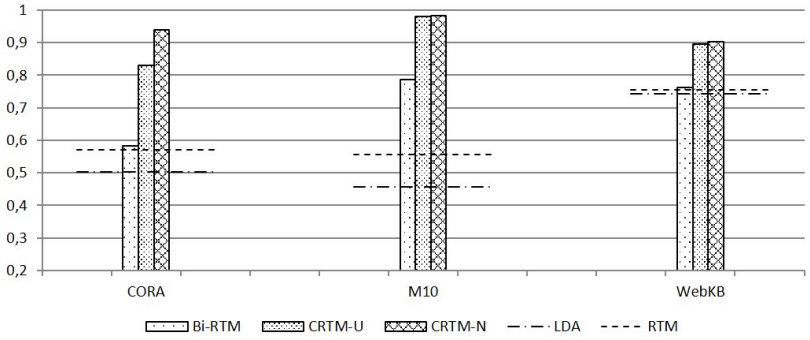

Figure 4: Micro-F1 measure of the models across all the datasets. The plot shows the best performance of the average behavior of the models, considering different percentages of constraints introduced in the training phase and in the test phase.

counterpart CRTM-U, because it can handle the documents' length issue of the Cora dataset. Bi-RTM outperforms standard RTM and LDA, however in Cora and WebKB the improvement in the performance is small, meaning that modeling the document constraints using the link likelihood function $\psi_{\sigma'}$ may not be a promising solution.

We do not report LLDA in this evaluation, because, in LLDA, all of the words of a labeled document are associated with the same topic. This has the trivial effect of automatically label each document affected by a constraint in the test set with the correct class. CRTMs still have very promising results, and they can be applied in more realistic cases, i.e. when we do not know the exact labels of documents, but we know that two documents belong to the same class. In this scenario, LLDA cannot be used.

A further comparison is shown in Figure 5, where the results are reported in terms of macro-F1 measure, with knowledge introduced both in training and testing. We can easily notice that macro-F1 values are lower than the micro-F1 ones, highlighting that all the models are negatively affected by the class/topic size. This means that all the LDA-based models tend in general to fit better those classes with higher cardinality at the expenses of the minority classes. A set of additional results in terms of macro-F1 are reported in Appendix A.

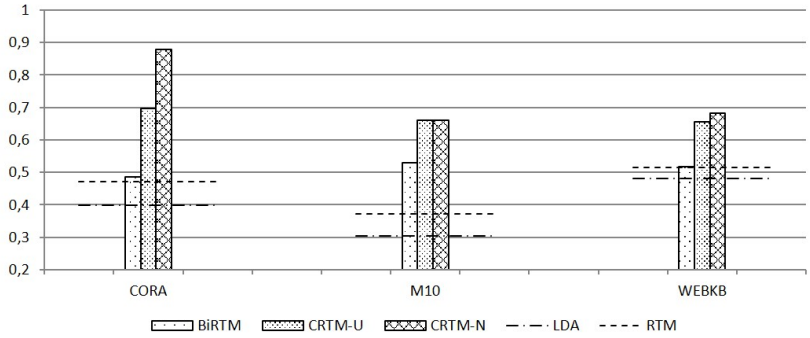

Figure 5: Macro-F1 measure of the models across all the datasets. The plot shows the best performance of the average behavior of the models, considering different percentages of constraints introduced in the training phase and in the test phase.

This behavior is mainly motivated by the symmetric and positive (> 1) values of the hyper-parameter α that regulates the corresponding document-topic distribution θ. In fact, this setting implies to have the same prior distribution of topics (and classes) for each document, originating therefore a posterior topic/classes distribution that is almost uniform and consequently balanced among different classes.

Even if CRTM is sensitive to the hyper-parameter α, it still outperforms the other baselines. The promising performance in terms of macro-F1 is mainly due to its abilities to smooth the posterior topic distributions by the introduction of constraints.

In order to show the complexity of the network obtained by the combination of links and must-constraints, we illustrate an example of the Cora benchmark. Figure 6 shows an instance of a document network when 0.2% of knowledge is introduced during the training phase. In particular, each node denotes a document, whose color represents the actual document class. The edges denote either a citation link or a must-constraint. Since must-constraints are allowed only between documents of the same class, this type of relationships form seven connected components that are visible observing the network. On the other hand, citations can exist either between same-class documents or documents

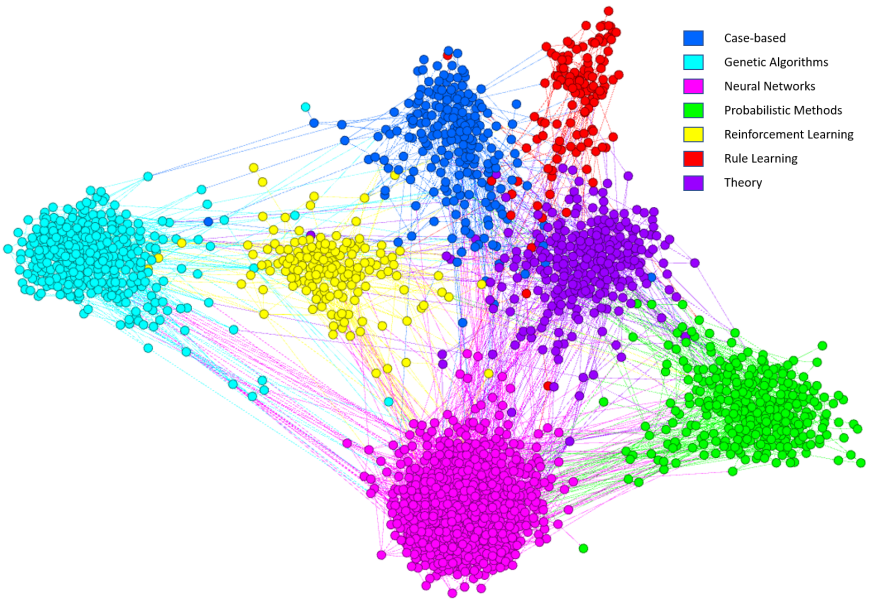

Figure 6: An example of the Cora network used during the training phase. Edges represent either citation links or must-constraint, where 0.2% of knowledge is incorporated. Nodes are colored with respect to their actual class.

belonging to different classes.

The density of the citation network together with the density of the constraints can have an impact on the classifier's performance, that decreases when too much knowledge is introduced. To better clarify this issue, we consider the ego network of the document 40886 (where 40886 is the original identifier of the document in the Cora dataset), as illustrated in Figure 7. In particular, the color of the node represents the actual class of a document, while color of the outline denotes the predicted class (e.g. documents 40886 and 429805 are classified as belonging to the class "Neural Network", but the first is correctly predicted while the second is misclassified).

As expected, the proposed model CRTM-N encourages all the purple nodes to have similar topic distributions and the classifier correctly predicts that all the documents belong to the class "Neural Network". Analogously, all the blue nodes are encouraged by both the must-constraints and the citation links to

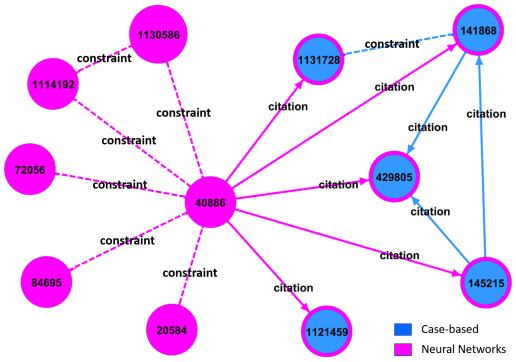

Figure 7: Ego network of document 40886 of the Cora dataset. The nodes are labeled by the original identifiers of the dataset and edges are labeled by the relationship type (citation are denoted by a straight line and must-constraint by a dashed line). The node color represents the actual class of a document, and the color of the outline denotes the predicted class.

have similar topic distributions and are assigned to the *same* class, though the predicted class is not correct. This error is likely due to the presence of citation links between documents of different classes (e.g. the citation between documents 40886 and 429805) combined with the must-constraints. If a document is misclassified, a must-constraint may propagate this error to all the other documents that are must-constrained to the misclassified one.

This error could be reduced through the use of cannot-constraints, that can be incorporated if two documents belong to different classes. In this way, a cannot-constraint between two documents would allow the topic distributions to be dissimilar, originating therefore a correct classifier's prediction.

7. Conclusions and Future Work

In this paper, we proposed Constrained Relational Topic Model (CRTM), a class of relational topic models that are able to incorporate some domain knowledge during the inference and training phases in the form of constraints. We also defined two document-constraint potential functions, and we investigated the models' performance in a document classification task by considering different quantities of prior knowledge. Experimental results on several relational

datasets have demonstrated the advantages of CRTM, using the proposed potential functions. As a future development, since the model depends on a set of parameters, grid search and Bayesian optimization techniques [4, 41] can be investigated to derive an optimal parameter configuration of CRTM. Finally, we propose to investigate the issue related to links between documents that belong to different classes by incorporating the cannot-constraints in the experimental results.

References

[1] A. Ahmed, L. Hong, and A. J. Smola. Nested Chinese Restaurant Franchise Process: Applications to User Tracking and Document Modeling. In *Proceedings of the 30th International Conference on Machine Learning, ICML*, pages 1426–1434, 2013.

[2] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 25–32, 2009.

[3] D. Andrzejewski, X. Zhu, M. Craven, and B. Recht. A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation Using First-Order Logic. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1171–1177, 2011.

[4] A. U. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On Smoothing and Inference for Topic Models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 27–34, 2009.

[5] H. Bai, Z. Chen, M. R. Lyu, I. King, and Z. Xu. Neural Relational Topic Models for Scientific Article Analysis. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018*, pages 27–36, 2018.

[6] S. Basu, M. Bilenko, and R. J. Mooney. A Probabilistic Framework for Semi-Supervised Clustering. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 59–68, 2004.

[7] K. Batmanghelich, A. Saeedi, K. Narasimhan, and S. Gershman. Nonparametric Spherical Topic Modeling with Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, page 537, 2016.

[8] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. *J. ACM*, 57(2):7:1–7:30, 2010.

[9] D. M. Blei and J. D. Lafferty. Correlated Topic Models. In *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, pages 147–154, 2005.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.

[11] J. L. Boyd-Graber and D. M. Blei. Syntactic Topic Models. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 185–192, 2008.

[12] J. Chang and D. M. Blei. Relational Topic Models for Document Networks. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 81–88, 2009.

[13] N. Chen, J. Zhu, F. Xia, and B. Zhang. Generalized Relational Topic Models with Data Augmentation. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1273–1279, 2013.

[14] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Discovering Coherent Topics using General Knowledge. In *Proceedings of*

the 22nd ACM International Conference on Information and Knowledge Management (CIKM), pages 209–218, 2013.

[15] Y. Cong, B. Chen, H. Liu, and M. Zhou. Deep Latent Dirichlet Allocation with Topic-Layer-Adaptive Stochastic Gradient Riemannian MCMC. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 864–873, 2017.

[16] R. Das, M. Zaheer, and C. Dyer. Gaussian LDA for Topic Models with Word Embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 795–804, 2015.

[17] G. Fei, Z. Chen, and B. Liu. Review Topic Discovery with Phrases using the Pólya Urn Model. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 667–676, 2014.

[18] Z. Gan, C. Chen, R. Henao, D. E. Carlson, and L. Carin. Scalable Deep Poisson Factor Analysis for Topic Modeling. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*, pages 1823–1832, 2015.

[19] T. L. Griffiths and M. Steyvers. Finding Scientific Topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, 2004.

[20] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating Topics and Syntax. In *Advances in Neural Information Processing Systems 18 (NIPS 2004)*, pages 537–544, 2004.

[21] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum. Topics in Semantic Representation. *Psychological review*, 114(2):211, 2007.

[22] A. Gruber, Y. Weiss, and M. Rosen-Zvi. Hidden Topic Markov Models. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 163–170, 2007.

[23] W. Guo, S. Wu, L. Wang, and T. Tan. Social-Relational Topic Model for Social Networks. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1731–1734, 2015.

[24] Y. He, C. Wang, and C. Jiang. Modeling Document Networks with Tree-Averaged Copula Regularization. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 691–699, 2017.

[25] A. Hefny, G. Gordon, and K. Sycara. Random Walk Features for Network-aware Topic Models. In *NIPS 2013 Workshop on Frontiers of Network Analysis*, volume 6, 2013.

[26] R. Henao, Z. Gan, J. Lu, and L. Carin. Deep Poisson Factor Modeling. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pages 2800–2808, 2015.

[27] J. H. Kim, D. Kim, S. Kim, and A. H. Oh. Modeling Topic Hierarchies with the Recursive Chinese Restaurant Process. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12*, pages 783–792, 2012.

[28] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma. Topic Modeling for Short Texts with Auxiliary Word Embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*, pages 165–174, 2016.

[29] W. Li and A. McCallum. Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations. In *Proceedings of the 23rd International Conference (ICML)*, pages 577–584, 2006.

[30] K. W. Lim and W. L. Buntine. Bibliographic Analysis with the Citation

Network Topic Model. In *Proceedings of the Sixth Asian Conference on Machine Learning (ACML)*, 2014.

[31] R. V. Lindsey, W. Headden, and M. Stipicevic. A Phrase-Discovering Topic Model Using Hierarchical Pitman-Yor Processes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 214–222, 2012.

[32] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. *J. Artif. Intell. Res.*, 30:249–272, 2007.

[33] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic Modeling with Network Regularization. In *Proceedings of the 17th International Conference on World Wide Web (WWW)*, pages 101–110, 2008.

[34] D. M. Mimno and A. McCallum. Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression. In *UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*, pages 411–418, 2008.

[35] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson. Improving Topic Models with Latent Feature Word Representations. *TACL*, 3:299–313, 2015.

[36] J. W. Paisley, C. Wang, D. M. Blei, and M. I. Jordan. Nested Hierarchical Dirichlet Processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):256–270, 2015.

[37] J. Petterson, A. J. Smola, T. S. Caetano, W. L. Buntine, and S. M. Narayanamurthy. Word Features for Latent Dirichlet Allocation. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1921–1929, 2010.

[38] Y. Tam and T. Schultz. Correlated Latent Semantic Model for Unsupervised LM Adaptation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 41–44, 2007.

[39] M. Wahabzada, Z. Xu, and K. Kersting. Topic Models Conditioned on Relations. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010*, pages 402–417, 2010.

[40] H. M. Wallach. Topic Modeling: Beyond Bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 977–984, 2006.

[41] H. M. Wallach, D. M. Mimno, and A. McCallum. Rethinking LDA: Why Priors Matter. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1973–1981, 2009.

[42] H. Wang, X. Shi, and D. Yeung. Relational Deep Learning: A Deep Latent Variable Model for Link Prediction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2688–2694, 2017.

[43] X. Wang, A. McCallum, and X. Wei. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM)*, pages 697–702, 2007.

[44] W. Yang, J. L. Boyd-Graber, and P. Resnik. Birds of a Feather Linked Together: A Discriminative Topic Model using Link-based Priors. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 261–266, 2015.

[45] W. Yang, J. L. Boyd-Graber, and P. Resnik. A Discriminative Topic Model using Document Network Structure. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 686–696, 2016.

[46] Y. Yang, D. Downey, and J. Boyd-Graber. Efficient Methods for Incorporating Knowledge into Topic Models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 308–317, 2015.

[47] A. Zhang, J. Zhu, and B. Zhang. Sparse Relational Topic Models for Document Networks. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML)*, pages 670–685, 2013.

[48] H. Zhao, L. Du, and W. L. Buntine. A Word Embeddings Informed Focused Topic Model. In *Proceedings of The 9th Asian Conference on Machine Learning, (ACML)*, pages 423–438, 2017.

[49] M. Zhou, Y. Cong, and B. Chen. Augmentable Gamma Belief Networks. *Journal of Machine Learning Research*, 17:163:1–163:44, 2016.

[50] Y. Zhu, X. Yan, L. Getoor, and C. Moore. Scalable Text and Link Analysis with Mixed-Topic Link Models. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 473–481, 2013.

Appendix A.

In the following, Figures A.8, A.9, and A.10 show some additional results on the three benchmark datasets, measured in terms of macro-F1 score. Knowledge is introduced only in the training phase, with different percentages.

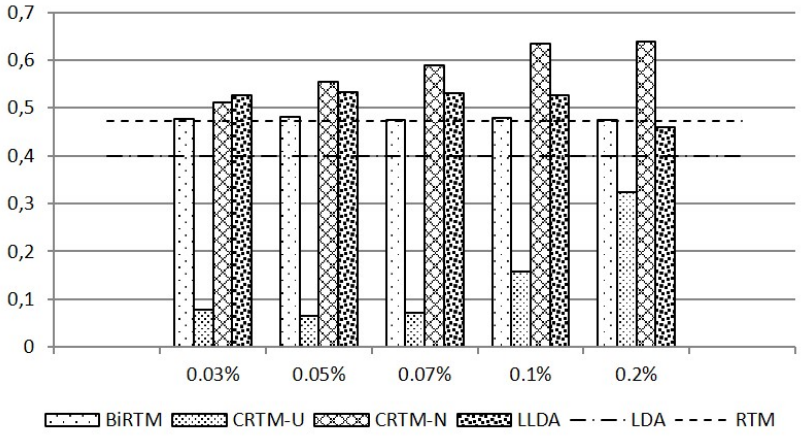

Figure A.8: Macro-F1 performance of the compared models on Cora.

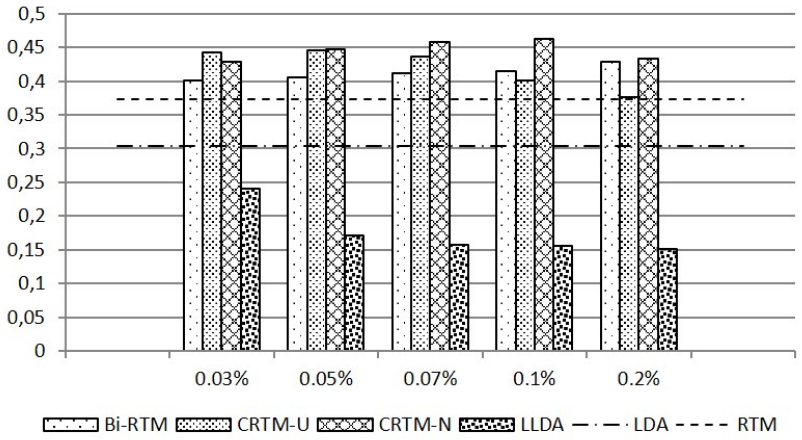

Figure A.9: Macro-F1 performance of the compared models on M10.

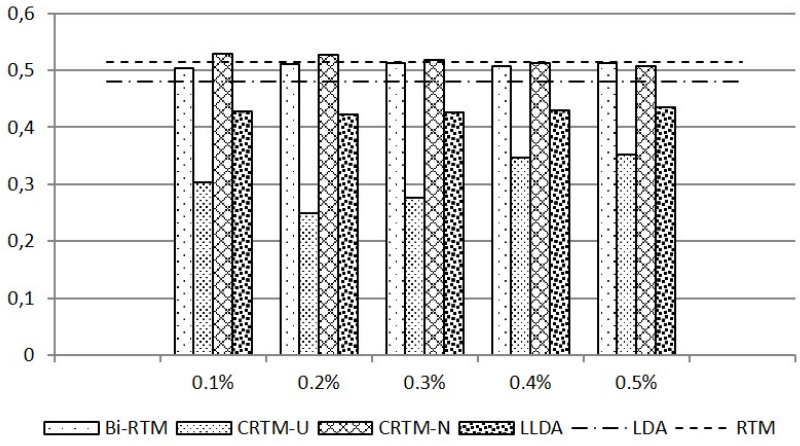

Figure A.10: Macro-F1 performance of the compared models on WebKB.