Which Matters Most? Comparing the Impact of Concept and Document Relationships in Topic Models

Silvia Terragni[♣], **Debora Nozza**[♠], **Elisabetta Fersini**[♣], **Enza Messina**[♣]
[♣]University of Milano-Bicocca, Milan,
[♠]Bocconi University, Milan
[♣] s.terragni4@campus.unimib.it, [♠]debora.nozza@unibocconi.it,
[♣]{elisabetta.fersini, enza.messina}@unimib.it

Abstract

Topic models have been widely used to discover hidden topics in a collection of documents. In this paper, we propose to investigate the role of two different types of relational information, i.e. document relationships and concept relationships. While exploiting the document network significantly improves topic coherence, the introduction of concepts and their relationships does not influence the results both quantitatively and qualitatively.

1 Introduction

Topic models are a suite of generative probabilistic models aimed at discovering thematic information (or topics) of an unstructured collection of documents. These models, including the well-known Latent Dirichlet Allocation (LDA) (Blei et al., 2003), usually consider texts as the unique source of information and are based on the assumption that texts are independent and identically distributed (i.i.d. assumption). However, in several real-world cases, documents are often characterized by an underlying relational structure: scientific papers can be related through citations, web pages can present hyperlinks between each other, and users in social networks can be friends. One of the first approaches that explicitly models the relationships between documents is Relational Topic Model (RTM) (Chang and Blei, 2009), based on the intuition that connected documents likely discuss the same topics.

Traditional topic models also assume that the topic assignment of a word is independent of other hidden topics, given the document's topic distribution. However, previous work proved that the introduction of additional knowledge about the relationships between words improves the coherence of the discovered topics (Yang et al., 2015b; Chen et al., 2013b,c). This type of relationship is commonly viewed as related to the concept of synonym, but this is not always the case in a real-world scenario because of word ambiguity. Following this intuition, it is thus important to take into consideration the concept behind the word alongside the word itself for understanding its relationship with other words, because it would permit to associate the same topic to words that are actually related and not only synonyms. For example, it would be possible to grasp that the word "engine", when associated with the concept of "search engine", is distant from "motor", but similar to "information retrieval". Few works investigate the use of named entities in topic models (Kim et al., 2012; Wang et al., 2017; Allahyari and Kochut, 2016), but none of them addresses the problem in relational settings.

Contribution In this paper, we investigate the role of two different types of relational information: (1) concept relationships between words and named entities obtained by Word Embeddings and (2) document-level relationships extracted by a document network. The impact of these two types of relational information is evaluated by considering traditional topic models and by introducing two novel Entity Constrained Topic Models. The source code has been made available at the following link: https://github.com/MIND-Lab/EC-RTM.

2 Related Work

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a generative probabilistic model that describes a document corpus through a set of topics K, seen as distributions of words over a fixed vocabulary. A document is assumed as composed of a mixture of the topics, following a Dirichlet distribution. Words are generated according to the topics drawn from this mixture. LDA can be extended by considering different types of relational information.

Proceedings of the First Workshop on Insights from Negative Results in NLP, pages 32–40

ISBN 978-1-952148-66-8

Word-level Relational Topic Models relax the independence assumption of words in a document or in a topic. They can be roughly divided into models that encode word-order (Wang et al., 2007; Gruber et al., 2007; Lindsey et al., 2012; Fei et al., 2014; Wallach, 2006) and syntactic dependencies (Griffiths et al., 2004; Boyd-Graber and Blei, 2008), and models that incorporate semantic or domain knowledge relationships (Andrzejewski et al., 2009, 2011; Chen et al., 2013b; Yang et al., 2015b). Lately, the growing interest in word embeddings has led to the incorporation of the relationships deriving from word embeddings (Petterson et al., 2010; Zhao et al., 2017; Das et al., 2015; Nguyen et al., 2015; Li et al., 2016; Batmanghelich et al., 2016; Nozza et al., 2016).

Document-level Relational Topic Models assume that two linked documents are more likely to have similar topic distributions. Relational Topic Model (RTM) and its extensions (Chen et al., 2013a; Terragni et al., 2020; Zhang et al., 2013; Yang et al., 2015a, 2016), ground on LDA and model each link as a binary variable considering the existence of a link between pairs of documents. Other approaches include the regularized topic models (He et al., 2017; Mei et al., 2008), which augment the model's objective function with a network regularization penalty, and the Dirichlet Multinomial Regression (Mimno and McCallum, 2008) and its extensions (Hefny et al., 2013; Wahabzada et al., 2010), incorporating links by viewing them as per-document attributes. A promising paradigm uses neural variational inference to infer topics (Miao et al., 2016; Bianchi et al., 2020a,b). Neural Relational Topic Model (NRTM) (Bai et al., 2018), is based on Stacked Variational AutoEncoder (SVAE) to infer topics and predict links using a multilayer perceptron.

3 Entity Constrained Topic Models

We propose **Entity Constrained Latent Dirichlet Allocation (EC-LDA)** and **Entity Constrained Relational Topic Models (EC-RTM)**, two classes of models aimed at incorporating entity-entity and entity-word relationships in traditional topic models. Following (Yang et al., 2015b; Terragni et al., 2020), we constrain the joint distribution of LDA and RTM through the use of potential functions that model entity-entity and/or entity-word relationships. The potential can be factored out of the joint distribution and the posterior can be derived using a collapsed Gibbs sampling for inference. In addition to EC-LDA, EC-RTM also assumes that two linked documents are likely to discuss the same topics. We report the joint distributions of the proposed models in the Appendix A. For further details on Constrained Topic Models, we refer the reader to (Yang et al., 2015b; Terragni et al., 2020).

We define the vocabulary E containing the unique named entities of the corpus, and the vocabulary W containing the unique words. We derive the vocabulary Γ as the union of the word and named entity vocabularies. Relationships are denoted by the set of knowledge L and each piece of knowledge $l \in L$ is incorporated by a potential function $f_l(z, u)$, which represents a real-valued score for the hidden topic assignment z of the word or named entity token u.

We derive the knowledge L using Skip-Gram (Mikolov et al., 2013). Given a word embeddings training set composed of a large but finite set Λ, the word embeddings model can be expressed as a mapping function $C' : \Gamma \mapsto \mathbb{R}^t$. For each token $u \in \Gamma$, we define a *must-constraint* set L_u^m, containing words and named entities that are likely to share the same themes of u. L_u^m is defined as:

$$L_u^m = \{v \in \Gamma | sim(C'(u), C'(v)) > \epsilon_m\} \quad (1)$$

where sim is the cosine similarity between two vectors, and ϵ_m is a given threshold. We also define a *cannot-constraint* set L_u^c, that contains the words and named entities that are not likely to share the same themes of u. L_u^c is defined as:

$$L_u^c = \{v \in \Gamma | sim(C'(u), C'(v)) < \epsilon_c\} \quad (2)$$

where ϵ_c is a given threshold.

An example of a must-constraint set for the named entity "*Artificial neural network*" may be {*Artificial neuron*, *ANN*, *perceptron*} which contains named entities that are likely to be assigned to the same topic. Analogously, an example of cannot-constraint set for the named entity "*Artificial neural network*" may be {*Olympic games, Athlete*} which denotes named entities related to sports and not to Machine Learning.

3.1 Entity-Entity Potential Function

We specify an entity-entity potential function that models the relationships between named entities. Let $N_{ze'}$ be the maximum between 1 and the topic-entities counts, i.e. the number of occurrences of e'

assigned to topic z. The function $f_l(z,u)$ is as follows:

$$f_l(z,u) = \begin{cases} \sum\limits_{\substack{e' \in L_u^m \\ e' \in E}} \log N_{ze'} + \sum\limits_{\substack{e' \in L_u^c \\ e' \in E}} \log \frac{1}{N_{ze'}} & \text{if } u \in E \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The function increases the probability that the entity u will be assigned to the same topics as those of the entities belonging to L_u^m. Similarly, the potential function decreases the probability that a named entity u will be drawn from the same topics as those of entities contained in the L_u^c.

The models that can encode the Entity-Entity (EE) potential function will be referred to EC-LDA-EE and EC-RTM-EE.

3.2 Entity-Word Potential Function

Let $N_{zw'}$ be the maximum between 1 and the topic-word counts, i.e. the counts of word w' assigned to topic z. The following potential function deals with relationships between entities and word tokens:

$$f_l(z,u) = \begin{cases} \sum\limits_{\substack{w' \in L_u^m \\ w' \in W}} \log N_{zw'} + \sum\limits_{\substack{w' \in L_W^c \\ w' \in W}} \log \frac{1}{N_{zw'}} & \text{if } u \in E \\ \sum\limits_{\substack{e' \in L_u^m \\ e' \in E}} \log N_{ze'} + \sum\limits_{\substack{e' \in L_u^c \\ e' \in E}} \log \frac{1}{N_{ze'}} & \text{if } u \in W \end{cases} \quad (4)$$

The potential function models the following cases:

- if u is a named entity, then we consider only the words that are contained in u's must- and cannot-constraint sets, i.e. L_u^m and L_u^c;
- if u is a word, then we consider only the named entities that are contained in u's must- and cannot-constraint sets, i.e. L_u^m and L_u^c.

The models encoding Entity-Word (EW) relationships are named EC-LDA-EW and EC-RTM-EW.

4 Experimental setting

Datasets The experimental investigation has been performed on two relational benchmark datasets: (1) *Cora-ML* (McCallum et al., 2005), a citation network on the set of Machine Learning papers (Sen et al., 2008) and (2) *WebKB*[1], a website dataset collected from 4 different universities, where links are hyperlinks. Table 1 reports the basic statistics of the datasets.

[1] www.cs.cmu.edu/~WebKB/ILP-data.html

Datasets	#Docs	#Links	Document Type	Link Type
Cora-ML	2,708	5,278	Title+Abstract	Citation
WebKB	877	1,608	Webpage	Hyperlink

Table 1: Statistics of benchmark datasets.

Preprocessing The identification of named entities in text is typically performed through a series of techniques that refer to the task of Named Entity Recognition (NER) (Fersini et al., 2014; Ritter et al., 2011; Li et al., 2020). Once the named entities are recognized, the next step is to associate them to unambigous concepts, as for example resources in a Knowledge Base. This process is known as the task of Named Entity Linking (NEL) (Cucerzan, 2007; Dredze et al., 2010; Basile et al., 2015; Cecchini et al., 2016; Nozza et al., 2019).

In this paper, we used the DBPedia Spotlight tool (Mendes et al., 2011) (confidence = 0.5 and support = 0.0) to identify named entities in the text and associate them to DBPedia units. We added the prefix "NE/" to each identified entity to discriminate it from words. We applied a common preprocessing technique on the text. We considered only must-constraints, that have been extracted from Wikipedia2Vec (Yamada et al., 2018). For details on the hyperparameters and preprocessing, see the Appendix A.

Compared Models We compared the proposed models (i.e., EC-LDA-EE, EC-LDA-EW and EC-RTM-EE, EC-RTM-EW) with the significant state-of-the-art approaches, i.e. Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Relational Topic Model (RTM) (Chang and Blei, 2009), Stacked Variational Auto-Encoder (SVAE) and Neural Relational Topic Model (NRTM) (Bai et al., 2018).

Metrics We use *KL-U*, *KL-V*, and *KL-B* to measure semantic importance and identify junk and insignificant topics (AlSumait et al., 2009). We also measure how different are the topics from each other by computing Topic Diversity (***TD***) (Dieng et al., 2019). Finally, we consider two metrics of topic coherence, i.e. *NPMI* (Aletras and Stevenson, 2013) and C_V (Röder et al., 2015) that measure how much the 10-top words of a topic are related to each other. The scores are computed using the Palmetto toolkit[2] and Wikipedia[3] as reference corpus.

[2] http://github.com/dice-group/Palmetto

[3] English Wikipedia dump of the 23rd of May, 2019.

	KL–U			*KL–V*			*KL–B*			*TD*			*NPMI*			C_V		
	10	30	50	10	30	50	10	30	50	10	30	50	10	30	50	10	30	50
LDA	1.855	1.572	1.259	1.226	1.231	1.059	0.052	0.119	0.168	0.816	0.736	0.654	0.098	0.080	0.071	0.399	0.389	0.386
RTM	2.001	2.046	1.820	1.357	1.563	1.460	0.095	**0.207**	**0.283**	0.814	**0.747**	0.666	0.099	0.082	0.071	0.348	0.391	0.392
EC-LDA-EE	1.845	1.520	1.375	1.225	1.238	1.066	0.052	0.119	0.167	0.814	0.742	0.659	0.098	0.079	0.069	0.397	0.390	0.389
EC-LDA-EW	1.800	1.518	1.381	1.230	1.236	1.065	0.052	0.119	0.168	0.817	0.740	0.660	0.094	0.079	0.070	0.395	0.389	0.387
EC-RTM-EE	2.033	**2.082**	**1.849**	**1.362**	1.564	**1.472**	0.095	0.205	0.280	0.817	0.747	**0.675**	**0.099**	0.081	0.071	0.402	0.394	0.392
EC-RTM-EW	**2.079**	1.990	1.643	1.361	**1.565**	1.470	**0.096**	0.206	0.282	0.820	0.746	0.671	0.098	**0.082**	**0.072**	0.340	0.392	0.392
SVAE	-	-	-	-	-	-	-	-	-	**0.893**	0.694	0.577	-0.099	-0.095	-0.096	**0.456**	**0.456**	**0.453**
NRTM	-	-	-	-	-	-	-	-	-	0.857	0.525	0.381	-0.083	-0.082	-0.082	0.442	0.447	0.446

Table 2: Performance on the *Cora-ML* dataset with the number of topics equal to 10, 30, 50.

	KL–U			*KL–V*			*KL–B*			*TD*			*NPMI*			C_V		
	10	30	50	10	30	50	10	30	50	10	30	50	10	30	50	10	30	50
LDA	1.695	1.256	1.130	1.054	0.943	0.775	0.069	0.142	0.199	0.761	0.617	0.538	0.039	0.040	0.030	0.378	0.379	0.379
RTM	**1.986**	1.795	1.430	**1.202**	1.239	1.109	**0.119**	0.225	**0.303**	0.760	0.608	0.532	0.043	0.043	0.036	0.377	0.380	0.380
EC-LDA-EE	1.643	1.289	1.061	1.055	0.948	0.780	0.069	0.143	0.200	0.769	0.623	0.542	0.043	0.041	0.033	0.379	0.380	0.381
EC-LDA-EW	1.736	1.345	1.075	1.062	0.981	0.784	0.069	0.138	0.198	0.764	**0.651**	**0.547**	0.042	0.038	0.033	0.376	0.381	0.382
EC-RTM-EE	1.867	**1.944**	1.468	1.199	1.246	1.119	0.118	**0.226**	0.303	0.760	0.612	0.536	**0.048**	**0.043**	**0.039**	0.377	0.382	0.381
EC-RTM-EW	1.979	1.786	**1.646**	1.199	**1.294**	**1.127**	0.117	0.217	0.302	0.759	0.639	0.543	0.045	0.042	0.036	0.377	0.382	0.384
SVAE	-	-	-	-	-	-	-	-	-	**0.829**	0.563	0.454	-0.116	-0.110	-0.112	**0.460**	0.450	0.452
NRTM	-	-	-	-	-	-	-	-	-	0.734	0.360	0.283	-0.114	-0.117	-0.119	0.454	**0.455**	**0.458**

Table 3: Performance on the *WebKB* dataset with the number of topics equal to 10, 30, 50.

5 Experimental Results

Quantitative Results Tables 2 and 3 show the performance of the models in terms of all the considered scores over an increasing number of topics on the datasets.[4] Results show that models that consider relational information generally obtain higher performance than their non-relational counterpart. Differently, the introduction of the concept constraints in EC-RTM-EE and EC-RTM-EW models does not seem to provide significant improvements with respect to RTM. This can be motivated by the fact that the constraint sets additionally included in the EC-RTM models are already captured in the word-topic distribution obtained by RTM.
Different behaviors can be observed for the C_V scores, for which NRTM and SVAE obtain significantly higher performance. This opposite trend with respect to the other topic scores can be explained by the fact that C_V rewards the presence of rare words even if they are contained in junk topics as stated by the author of (Röder et al., 2015)[5].

Qualitative Results In Table 4, we show the top-10 words for Cora-ML concerning an example topic "Genetic Programming" for EC-RTM-EE, EC-RTM-EW, LDA, RTM, SVAE, and NRTM. To analyze if the named entity annotation can contribute to topic interpretability, we report the words of LDA and RTM (referred as LDA* and RTM*) run on Cora-ML composed of words only. As expected from the quantitative results, the topics extracted by the proposed models do not significantly differ from RTM*, further demonstrating the hypothesis that the imposed constraints were already captured by the original model.
Qualitative considerations can be made regarding the exploitation of the novel entity-level modeling of the documents. While this representation leads to topics containing explicit concepts (e.g., "NE/Genetic_programming"), topics obtained by RTM* seem to be equally interpretable because they can identify named entities in the form of

Models	**Top-10 words**
LDA*	problem genetic algorithms problems programming search optimization fitness population space
RTM*	genetic control programming fitness reinforcement population algorithms paper environment behavior
EC-RTM-EE	NE/Genetic_programming programs NE/Genetic_algorithm population fitness genetic evolutionary program NE/Evolution strategies
EC-RTM-EW	NE/Genetic_programming NE/Genetic_algorithm population fitness genetic evolutionary NE/Evolution encoding operator operators
SVAE	koza NE/Multidisciplinary_design_optimization splice bitsback NE/Genetic_programming fitness orientation NE/Ploidy NE/Exon coded
NRTM	genetic reactive NE/Genetic_programming NE/Case casebased neuroevolution ssa NE/Genetic_algorithm coevolutionary problemsolving

Table 4: "Genetic Programming" topic in *Cora-ML*.

[4] Computing the KL- metrics is impractical for SVAE and NRTM since they do not model word- and document-topic distributions.

[5] https://bit.ly/3jApSAC

distinct words (e.g., "genetic, programming, algorithm"). Moreover, the difference in representation is only evident when named entities are composed of two or more words (e.g., "NE/Evolution" and "evolution" are equivalent). The benefit of applying NEEL techniques for recognizing named entities in topics may come in handy for automatically providing links to KB (such as Wikipedia), at the computational cost of discovering named entities. Moreover, the proposed novel potential function would allow users to artificially manipulate the model to derive explanations for the topic assignments or force entities in the same topic based on human domain knowledge.

Regarding SVAE and NRTM, their topics seem hard to interpret from a qualitative perspective, confirming the results of the quantitative evaluation.

6 Conclusion

We propose two classes of Entity Constrained Topic Models for incorporating different types of relational information. Results demonstrated that models exploiting document-level relationships achieve improvements with respect to their non-relational counterparts. Differently, concept relationships do not significantly improve either topic coherence or interpretability. As future work, we plan to investigate multi-relational topic models extracting other relationships from the data and to exploit contextual encoding method for entity representation also in multilingual settings (Devlin et al., 2019; Nozza et al., 2020).

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics, IWCS 2013, March 19-22, 2013, University of Potsdam, Potsdam, Germany*, pages 13–22. The Association for Computer Linguistics.

Mehdi Allahyari and Krys Kochut. 2016. Discovering coherent topics with entity topic models. In *Proceedings of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2016, Omaha, NE, USA, October 13-16, 2016*, pages 26–33.

Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. Topic significance ranking of LDA generative models. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009*, pages 67–82.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 25–32.

David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. 2011. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 1171–1177.

Haoli Bai, Zhuangbin Chen, Michael R. Lyu, Irwin King, and Zenglin Xu. 2018. Neural relational topic models for scientific article analysis. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 27–36.

Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro, and Fedelucio Narducci. 2015. UNIBA: Exploiting a Distributional Semantic Model for Disambiguating and Linking Entities in Tweets. In *Proc. of the 5th Workshop on Making Sense of Microposts co-located with the 24th International World Wide Web Conference*, volume 1395, page 62.

Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Samuel Gershman. 2016. Nonparametric spherical topic modeling with word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2020b. Cross-lingual contextualized topic models with zero-shot learning. *arXiv preprint arXiv:2004.07737*.

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

Jordan L. Boyd-Graber and David M. Blei. 2008. Syntactic topic models. In *Advances in Neural Information Processing Systems 21, Proceedings of the 22nd Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 185–192.

Flavio Massimiliano Cecchini, Elisabetta Fersini, Pikakshi Manchanda, Enza Messina, Debora Nozza, Matteo Palmonari, and Cezar Sas. 2016. UNIMIB@NEEL-IT: Named Entity Recognition and Linking of Italian Tweets. In *Proc. of 3rd Italian Conference on Computational Linguistics & 5th*

Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, volume 1749.

Jonathan Chang and David M. Blei. 2009. Relational topic models for document networks. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*, pages 81–88.

Ning Chen, Jun Zhu, Fei Xia, and Bo Zhang. 2013a. Generalized relational topic models with data augmentation. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence, IJCAI 2013, Beijing, China, August 3-9, 2013*, pages 1273–1279.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malú Castellanos, and Riddhiman Ghosh. 2013b. Discovering coherent topics using general knowledge. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 209–218.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malú Castellanos, and Riddhiman Ghosh. 2013c. Leveraging Multi-Domain Prior Knowledge in Topic Models. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2071–2077.

Silviu Cucerzan. 2007. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. 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, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 795–804.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019. Topic modeling in embedding spaces. *CoRR*, abs/1907.04907.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Proc. of the 23rd International Conference on Computational Linguistics*, pages 277–285.

Geli Fei, Zhiyuan Chen, and Bing Liu. 2014. Review topic discovery with phrases using the pólya urn model. In *Proceedings of the 25th International Conference on Computational Linguistics, COLING 2014, August 23-29, 2014, Dublin, Ireland*, pages 667–676.

Elisabetta Fersini, Enza Messina, Giovanni Felici, and Dan Roth. 2014. Soft-constrained inference for Named Entity Recognition. *Information Processing & Management*, 50(5):807–819.

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

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2004. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 537–544.

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March 21-24, 2007*, pages 163–170.

Yuan He, Cheng Wang, and Changjun Jiang. 2017. Modeling document networks with tree-averaged copula regularization. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017*, pages 691–699.

Ahmed Hefny, Geoffrey Gordon, and Katia Sycara. 2013. Random walk features for network-aware topic models. In *NIPS 2013 Workshop on Frontiers of Network Analysis*, volume 6.

Hyungsul Kim, Yizhou Sun, Julia Hockenmaier, and Jiawei Han. 2012. ETM: entity topic models for mining documents associated with entities. In *Proceedings of the 12th IEEE International Conference on Data Mining, ICDM 2012*, pages 349–358.

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. 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 2016, Pisa, Italy, July 17-21, 2016*, pages 165–174.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

Robert V. Lindsey, William Headden, and Michael Stipicevic. 2012. 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 2012, July 12-14, 2012, Jeju Island, Korea*, pages 214–222.

Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. 2005. Topic and role discovery in social networks. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI-05, Edinburgh, Scotland, UK, July 30 - August 5, 2005*, pages 786–791.

Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. 2008. Topic modeling with network regularization. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*, pages 101–110.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011*, ACM International Conference Proceeding Series, pages 1–8. ACM.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1727–1736. JMLR.org.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

David M. Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9-12, 2008*, pages 411–418.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2016. Unsupervised irony detection: a probabilistic model with word embeddings. In *International Conference on Knowledge Discovery and Information Retrieval*, volume 2, pages 68–76. SCITEPRESS.

Debora Nozza, Cezar Sas, Elisabetta Fersini, and Enza Messina. 2019. Word embeddings for unsupervised named entity linking. In *International Conference on Knowledge Science, Engineering and Management*, pages 115–132. Springer.

James Petterson, Alexander J. Smola, Tibério S. Caetano, Wray L. Buntine, and Shravan M. Narayanamurthy. 2010. Word features for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, pages 1921–1929.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015*, pages 399–408.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. 2008. Collective classification in network data. *AI Magazine*, 29(3):93–106.

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2020. Constrained relational topic models. *Information Sciences*, 512:581 – 594.

Mirwaes Wahabzada, Zhao Xu, and Kristian Kersting. 2010. Topic models conditioned on relations. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III*, pages 402–417.

Hanna M. Wallach. 2006. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning, (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, pages 977–984.

Qilin Wang, Dandan Song, and Xiuquan Li. 2017. Incorporating entity correlation knowledge into topic modeling. In *Proceedings of the IEEE International Conference on Big Knowledge, ICBK 2017, Hefei, China, August 9-10, 2017*, pages 254–258.

Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. 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 2007), October 28-31, 2007, Omaha, Nebraska, USA*, pages 697–702.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2018. Wikipedia2vec: an optimized tool for learning embeddings of words and entities from wikipedia. *arXiv preprint arXiv:1812.06280*.

Weiwei Yang, Jordan L. Boyd-Graber, and Philip Resnik. 2015a. 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 2015, Lisbon, Portugal, September 17-21, 2015*, pages 261–266.

Weiwei Yang, Jordan L. Boyd-Graber, and Philip Resnik. 2016. A discriminative topic model using document network structure. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Yi Yang, Doug Downey, and Jordan L. Boyd-Graber. 2015b. Efficient methods for incorporating knowledge into topic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 308–317.

Aonan Zhang, Jun Zhu, and Bo Zhang. 2013. Sparse relational topic models for document networks. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part I*, pages 670–685.

He Zhao, Lan Du, and Wray L. Buntine. 2017. A word embeddings informed focused topic model. In *Proceedings of The 9th Asian Conference on Machine Learning, ACML 2017, Seoul, Korea, November 15-17, 2017.*, pages 423–438.

A Appendix

A.1 Preprocessing

We lowercased the text, removed English stopwords and words occurring less than 10 times, and filtered out documents composed of less than 2 words. Details on the vocabulary composition are reported in Table 5.

A.2 Hyperparameters

Each experiment, with a given set of parameters, is repeated for 100 times and the performance measures are averaged by the number of the samples.

The hyperparameters α and β are set equal to $50/K$ and 0.1 respectively (as reported in (Griffiths and Steyvers, 2004)) for all the considered models. All the compared models are trained for 1,500 Gibbs iterations.

In our evaluation, we consider only must-constraint relations that can be generated by entities and words. To select the most appropriate value for the threshold ϵ_m, we studied the performance of the topic coherence of our models by varying the value of the parameter. The values for the models with the potential functions EE and EW are, respectively, 0.8 and 0.7 for the dataset Cora, and 0.6 and 0.6 for WebKB.

A.3 Joint Distributions of the Proposed Models

For the sake of completeness, we report the joint distribution of the proposed models. Entity Constrained Latent Dirichlet Allocation (EC-LDA) defines the following joint probability distribution:

$$P(\mathbf{u}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\Phi} | \alpha, \beta, L) \propto \tag{5a}$$

$$\prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(u_{nd}|\Phi_{z_{nd}}) p(z_{nd}|\theta_d) \tag{5b}$$

$$\prod_{k}^{K} p(\Phi_k|\beta) \cdot \xi(\mathbf{z}, L) \tag{5c}$$

where

- D denotes the set of documents
- N_d is the length of document d
- K denotes the fixed number of topics
- $\mathbf{u}$ denotes the set of word and named entity tokens
- $\mathbf{z}$ represents the set of topic assignments
- $\boldsymbol{\theta}$ represents the document-topic distribution
- $\boldsymbol{\Phi}$ denotes the topic-word distribution
- α and β are the Dirichlet hyperparameters related to θ and Φ
- $\xi(\mathbf{z}, L) = \prod_{z \in \mathbf{z}} \exp f_l(z, u)$.

Similarly, the joint probability distribution of Entity Constrained Relational Topic Models is defined as follows:

$$P(\mathbf{u}, \mathbf{z}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\Phi} | \alpha, \beta, \eta, \nu, L) \propto \tag{6a}$$

$$\prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(u_{nd}|\Phi_{z_{nd}}) p(z_{nd}|\theta_d) \tag{6b}$$

$$\prod_{k}^{K} p(\Phi_k|\beta) \prod_{\substack{d,d' \in D \\ d' \neq d}} \psi_\sigma(y_{d,d'}|z_d, z_{d'}, \eta, \nu) \cdot \xi(\mathbf{z}, L) \tag{6c}$$

where ψ_σ is the link probability function defined as $\psi_\sigma(y = 1) = \sigma(\eta^T(\overline{\mathbf{z}}_d \circ \overline{\mathbf{z}}_{d'}) + \nu)$, σ is the sigmoid function and $\overline{\mathbf{z}}_d = \frac{1}{N_d} \sum_n z_{nd}$. The link

	Processed corpus			Unprocessed corpus
	# unique entities	**# unique words**	**# unique entities and words**	**# unique words**
Cora	384	2,675	3,059	3,012
WebKB	355	1,874	2,229	2,247

Table 5: Summary of the vocabularies for the benchmark datasets before and after the preprocessing phase.

function models each per-pair binary variable related to links as a logistic regression (with hidden covariates), parameterized by coefficients η and intercept ν.

A.4 Computing Infrastructure

Experiments were run on three common computers using CPUs. Models can be run with basic infrastructure. Two computers have 8GB of RAM and the other has 16GB of RAM.