Enhancing Relational Topic Models with Named Entity Induced Links

Felix Kuhr, Mathis Lichtenberger, Tanya Braun and Ralf Möller
University of Lübeck
Institute of Information Systems
Ratzeburgerallee 160, 23562 Lübeck
{kuhr,lichtenberger,braun,moeller}@ifis.uni-luebeck.de

***Abstract*—Relational topic modeling as an extension to classical topic modeling assumes that documents with some form of link between the documents share topics. The links between documents are given from hyperlinks in web documents, citations in articles, or friendships in social networks. In this work, we consider links between documents induced from named entities: Two documents are linked to each other if both documents have a named entity in common. We present a case study on the performance of relational topic modeling using named-entity induced links between documents. Comparing the prediction accuracy with different sets of named-entity induced links, the results show that additional links between documents can increase the performance of topic models.**

I. Introduction

Latent Dirichlet allocation (LDA) allows for modeling topics of a corpus of documents, assuming documents are a mixture of topics, which determine the words in them [4]. Various topic modeling approaches have been developed that extend the original LDA. One set of models relaxes assumptions of LDA, e.g., relaxing the bag-of-words assumption s.t. the order of words is incorporated by generating words conditioned on the previous word or relaxing the assumption that the order of documents within the corpus is irrelevant, which accounts for topics changing over time [2]. Other topic modeling approaches enhance LDA by incorporating structure from metadata, e.g., the author-topic model [9].

In 2009, Chang and Blei have introduced the relational topic model (RTM) [3]. RTM considers links between documents in addition to their textual content to model topics of a corpus. The main assumption is that two linked documents share topics. Given an explicit link structure, RTM can show an increase in prediction accuracy compared to a topic model like LDA, which does not consider any links. The link structure for RTM may come from a citation network of a corpus of articles, from a network of hyperlinks in a corpus of web pages, or from a social network of friends in a corpus of postings.

In this work, we consider an implicit link structure available from extractable named entitys (NEs) shared between documents, instead of creating links from explicit metadata like citations or hyperlinks. We call our approach NE-RTM as we use RTM with implicit links via NEs . Over the last years, NEs have been used to improve the quality of discovered topics [6] or entity prediction [7]. The underlying assumption is that documents in a corpus share topics if documents share NEs, e.g., references to the same person, place, or company.

We collect two new data sets from the free online encyclopedia *Simple Wikipedia* and the *The New York Times* newspaper for a case study evaluating the performance of NE-RTM. Both data sets have an explicit link structure for RTM through hyperlinks as well as a plethora of NEs to induce links for NE-RTM. We analyze the attributes of NEs, their categories, and their perceived relevance to answer the question of which type of NEs should be considered for introducing links between documents within a corpus. Comparing the prediction accuracy of NE-RTM using different types of NE-induced links, the results show that additional links can increase the topic modeling performance. In summary, the contribution of this paper is two-fold, (i) analyzing the properties of NEs improving the performance of the topic model and (ii) comparing the performance between LDA, RTM, and NE-RTM with different settings for NE-induced links.

The remainder of this paper is structured as follows. Section II presents background information. Section III contains details about NE-induced links for RTM as well as results for the two data sets, showing the potential of an implicit link structure. Section IV concludes with future work.

II. Preliminaries

This section summarizes the main parts of RTM and named-entity recognition (NER). We describe only those parts that are of interest to follow our idea of NE-induced links for enhancing the performance of the RTM, i.e., NE-RTM.

RTM [3] is based on LDA [4], which describes a corpus $\mathcal{D}$ of documents using a set of K *topics*. The topics are represented as distributions over a fixed vocabulary $\mathcal{V}$ of all words in the documents of $\mathcal{D}$. The generative process of LDA is defined for each document as follows:

1. Choose a topic distribution $\theta_d \sim Dir(\alpha)$ for document $d \in \mathcal{D}$, where $Dir(\alpha)$ is a Dirichlet distribution from hyperparameter α.
2. Choose a word distribution $\phi_k \sim Dir(\beta)$ for a topic $k \in \{1, ..., K\}$, where $Dir(\beta)$ is a Dirichlet distribution from hyperparameter β.
3. For each word position $j \in \{1, ..., N\} \in d$:
   a) Choose a topic $z_{d,j} \sim Multinomial(\theta_d)$.
   b) Choose a word $\omega_{d,j} \sim Multinomial(\phi_{z_{d,j}})$.

The generative process of RTM consists of two parts. The first part is identical to the generative process of LDA. The second part concerns the links between documents, which are modeled as a binary variable, one variable for each pair of documents, and is defined as follows.

For each pair of documents $d, d' \in \mathcal{D}$:

$$\text{Draw binary link indicator: } y \mid z_d, z_{d'} \sim \psi(\cdot \mid z_d, z_{d'}),$$

where function ψ estimates the link probability between d and d' and is defined by:

$$\psi(y = 1) = \sigma(\boldsymbol{\eta}^T(\overline{\boldsymbol{z}}_d \circ \overline{\boldsymbol{z}}_{d'}) + \nu$$

where $\overline{z}_d = \frac{1}{N_d} \sum_n z_{d,n}$. The $\circ$ symbol denotes the Hadamad (element-wise) product. As in [3], function σ can be the sigmoid function or alternatively, the exponential mean function. If σ is the sigmoid function, it models each binary variable as a logistic regression with hidden covariates, parameterized by coefficients η and intercept ν. The covariates are constructed by the Hadamard product of $\overline{z}_d$ and $\overline{z}_{d'}$, modeling the similarity between the topics of d and d'. In case σ is the exponential mean function, the probabilities returned increase exponentially. RTM contains for each possible connection between all documents in $\mathcal{D}$ a link variable, and the number of link variables is given by $|\mathcal{D}|^2$. In summary, RTM defines a joint distribution over the words in each document and the links between them.

Methods of NER extract structured information from unstructured text. The goal of NER is identifying NEs and classifying them w.r.t. a predefined set of categories like names of persons, organizations, locations, etc. One of the first research approaches regarding NER uses heuristics and handcrafted rules to extract and recognize the names of companies.Nowadays, machine learning is usually the method of choice instead of relying only on handcrafted rules [11]. Over the years, many tools have been introduced for the task of NER, usually coming as part of an Information Extraction or natural language processing library.

III. Named-Entity Induced Links for RTM

We extend RTM by adding links between two documents if they have at least one NE in common. The assumption is that documents sharing NEs have similar topics since the same NEs appear in the text. We evaluate the performance using the perplexity measure on held-out data given by [5]

$$perplexity(w) = \exp\left(-\frac{\log\left(p(w)\right)}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)}}\right),$$

where $n^{(jd)}$ represents the occurrence of the j-th word in document d. The smaller the perplexity, the better the quality of the model. We fit the parameters in (NE-)RTM by performing Gibbs sampling, i.e., the word likelihood is given by:

$$\log\left(p(w)\right) = \sum_{d=1}^{|\mathcal{D}|}\sum_{j=1}^{|\mathcal{V}|} n^{(jd)} \log\left(\sum_{k=1}^{K} \theta_{d,k}\beta_{j,k}\right)$$

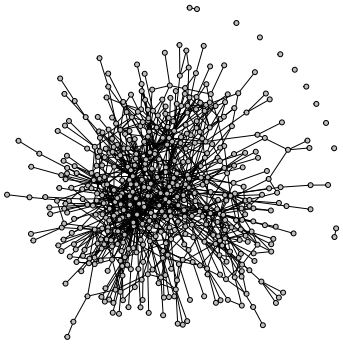

Fig. 1: Explicit network structure of *wiki-sparse-500* using hyperlinks between documents.

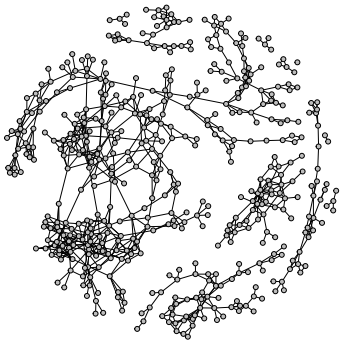

Fig. 2: Explicit network structure of *nyt-500* related coverage links between documents.

A. Data Sets

Data sets commonly used for evaluating topic models are *Cora* [1],*WebKB* [8], or the Proceedings of the National Academy of Science (PNAS), where citations or hyperlinks determine links. However, these data sets contain only very few NEs and are therefore not suitable for evaluating NE-RTM. Thus, we generate three new data sets to evaluate the performance: (i) *wiki-sparse-500*: 500 documents randomly selected from the English Simple Wikipedia. An average article consists of 348 words (241 without stop words). Each document has only few links to other documents. (ii) *wiki-dense-500*: 500 documents from the English Simple Wikipedia containing many documents about countries. An average article consists of 692 words (466 without stop words). Documents have a high number of links to other documents. (iii) *nyt-500*: 500 articles from The New York Times. An average article consists of 922 words (619 without stop words). Each document has links to articles containing related coverage leading to 942 links in total. After downloading all documents, we preprocess the documents by (i) eliminating stop words, (ii) performing word stemming, and (iii) tokenizing the text.

We use the Simple English Wikipedia, which aims to restrict itself to use simple words and grammar, aimed at children and adults who are learning the language. The use of simple language comes in handy for our purposes, as there are not as many words of low frequency in the data. For documents

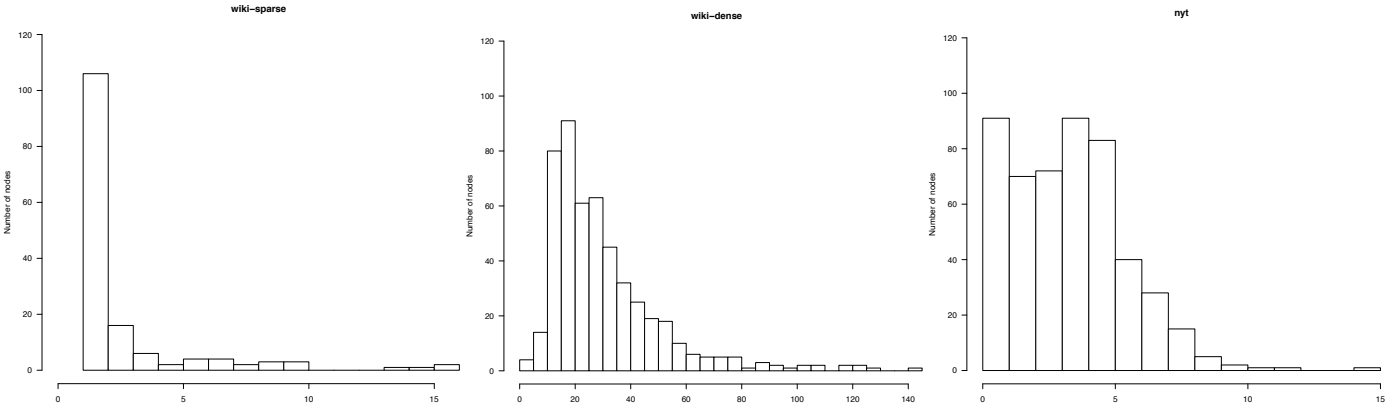

Fig. 3: Node degree histograms of the document network graphs for all three datasets.

TABLE I: Extractable NE categories

	wiki-sparse-500				*wiki-dense-500*				*nyt-500*			
#	category	total	bucket	links	category	total	bucket	links	category	total	bucket	links
1	Country	827	174	8717	Country	3006	257	55011	Country	1544	110	92854
2	IndustryTerm	664	486	1480	Continent	353	8	11403	Position	3406	1247	47732
3	Position	843	479	1285	Position	1128	458	7676	City	1191	274	35537
4	Continent	105	8	941	IndustryTerm	1139	691	6202	Organization	2712	974	26512
5	City	392	238	542	Organization	1401	897	4457	Person	3763	2134	18508
6	ProvinceOrState	205	91	353	Region	703	270	3569	Company	1681	815	9009
7	Organization	456	368	217	ProvinceOrState	358	149	2760	ProvinceOrState	724	101	6805
8	Technology	262	178	198	SportsEvent	235	119	2636	Continent	195	7	5155
9	MedicalCondition	275	219	112	City	1003	556	2452	IndustryTerm	1632	1057	3563
10	Company	316	275	109	NaturalFeature	851	476	1604	PublishedMedium	218	80	2400

in the *nyt-500* data set, there are only few hyperlinks between documents. But, there exists *related coverage* for most articles, containing links to up to five related articles, which we have used to generate the ground truth of links between documents.

Figures 1 and 2 represent the network structure for documents in the *wiki-sparse-500* and *nyt-500* dataset, respectively. Figure 3 contains the corresponding node degree histograms for the network structure of all three datasets.

B. Categories of NEs

We use Open Calais [10] to extract NEs from data sets and store the available relevance value and category for each NE. Open Calais provides many categories besides typical categories resulting in 36 different categories in the three datasets. Table I presents for all three data sets the top 10 NE categories (by links) containing the total number of occurrences, the number of unique names of that type (bucket), and the number of links that could be created from the corresponding NEs.

In the *wiki* data sets, categories *country*, *industry-term*, *position*, and *continent* lead to the most links between documents. The links from categories *organization* and *company* lead to only 10% of the number of links using categories *country*, *continent*, *position*, and *industry-term*. The *nyt-500* data set also features categories *country* and *position* as the top two categories. However, categories *continent* and *industry-term* are only on the 8th and 9th position, respectively.

We consider different sets of NE-induced links, (i) all categories (all), (ii) only *organization* and *company* (org), and (iii) all categories excluding *country*, *continent*, *position*, and *industry-term* (exc), to test the effect of the number of links and different categories introduced for each data set.

We also consider the following three thresholds for the relevance value r of NEs: (i) all NEs, ignoring the relevance value (all), (ii) only NEs having a relevance value ($r > 0$), and (iii) only NEs having a relevance value above 0.2 ($r > 0.2$).

C. Empirical Results

As RTM is a predictive model, we can measure the performance by withholding data in the training phase and examining the likelihood of predicting the held-out ground truth in a typical machine learning fashion. We generate the held-out set by removing half of the words from 10% of documents. After we have fitted the model, the topic probabilities for each document and the word distribution for each topic are known, which allows us to determine the predicted probability of the held-out data. For reasons of space, we only present results for $k = 15$, $\alpha = 0.1$, $\beta = 0.1$, and $\eta = 0.1$, having good performance over all three data sets. In Table II, we present the perplexities for all three data sets using the different settings for NE induced links for the relational topic model as described above. Additionally, we provide a baseline for the perplexity using the LDA topic model without any links (first row) and a basic RTM using only explicit links (second row). Explicit links are hyperlinks for Wikipedia documents and related coverage references for The New York Times articles. Rows 3 to 11 contain nine different NE-RTM

TABLE II: Topic model performance.

Topic Model	Links	Relevance	dense	sparse	nyt
LDA			1208.32	1156.91	1576.78
RTM			1196.16	1126.89	1573.71
NE-RTM	org	all	1193.89	1124.97	**1554.80**
NE-RTM	org	$r > 0$	1195.56	1125.47	1554.91
NE-RTM	org	$r > 0.2$	1195.26	*1127.19*	1570.68
NE-RTM	exc	all	1192.64	1121.04	*1625.36*
NE-RTM	exc	$r > 0$	1191.45	**1120.20**	*1621.90*
NE-RTM	exc	$r > 0.2$	1194.03	*1128.41*	1567.62
NE-RTM	all	all	1172.22	1207.44	*1824.66*
NE-RTM	all	$r > 0$	**1169.37**	1207.94	*1822.15*
NE-RTM	all	$r > 0.2$	1194.00	*1128.61*	*1576.14*

settings. The bold values show the best perplexity for each data set. The italic values represent a perplexity worse than the perplexity possible with basic RTM. The perplexity of RTM is better than the perplexity of LDA for all data sets. Additionally, as we have expected, the performance of NE-RTM is better than the perplexity of RTM using only explicit links. NE-induced links can enhance the performance of topic models, but simply adding links between documents sharing NEs does not automatically improve the performance and the choice of NE categories needs to be considered. Thus, it is no surprise that for all three data sets, the best perplexity is given by different configurations, since each data set represents a different characteristic. Overall, the data set with the most links, *wiki-dense-500*, benefits the most from added links through NEs. The best perplexity is given by considering all entities. The *nyt-500* data set mostly benefits from the (org) setup while (all) leads to worse results than basic RTM. One reason for this behaviour might be, that only articles about the same company contain similar text, while documents containing NEs like countries, continents, or positions appear often in different topics within a newspaper. The *wiki-sparse-500* data set benefits in 2 out of 3 cases but exhibits the smallest change in perplexity compared to the other two sets. The best perplexity is given by ignoring *countries*, *continents*, *positions*, and *industry-terms*. The *wiki-dense-500* data set shows the largest improvement. The data set which is linked to a high degree from the start seems to win the most by adding further links. The links induced by NEs from organizations increase the perplexity compared to RTM for all data sets except with the *wiki-sparse-500* data set and $r > 0.2$.

IV. Conclusion and Outlook

RTM uses explicit links between documents to improve the performance of a topic model for a corpus of documents. In this work, we add NE-induced links to further increase the performance. The results on three data sets show that the performance of RTM can be increased by adding links originating from NER. The performance depends on the documents in the corpus and the NEs used to induce links. Future work includes exploring more diverse data sets and analyzing the effect of different NE recognizers.

References

[1] Andrew McCallum, and Kamal Nigam, and Jason Rennie, and Kristie Seymore. "Automating the Construction of Internet Portals with Machine Learning". In: *Inf. Retr.* 3.2 (2000), pp. 127–163.

[2] David M. Blei and John D. Lafferty. "Dynamic topic models". In: *Machine Learning, Proc. of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*. 2006, pp. 113–120.

[3] Jonathan Chang and David M. Blei. "Relational Topic Models for Document Networks". In: *Proc. of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*. 2009, pp. 81–88.

[4] David M. Blei and Andrew Y. Ng and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[5] David Newman, and Arthur U. Asuncion, and Padhraic Smyth, and Max Welling. "Distributed Algorithms for Topic Models". In: *Journal of Machine Learning Research* 10 (2009), pp. 1801–1828.

[6] Katsiaryna Krasnashchok and Salim Jouili. "Improving Topic Quality by Promoting Named Entities in Topic Modeling". In: *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*. 2018, pp. 247–253.

[7] Linmei Hu, and Juan-Zi Li, and Zhihui Li, and Chao Shao, and Zhixing Li. "Incorporating Entities in News Topic Modeling". In: *Natural Language Processing and Chinese Computing - Second CCF Conference, NLPCC 2013, Chongqing, China, November 15-19, 2013, Proc.* 2013, pp. 139–150.

[8] Mark Craven, and Dan DiPasquo, and Dayne Freitag, and Andrew McCallum, and Tom M. Mitchell, and Kamal Nigam, and Seán Slattery. "Learning to Extract Symbolic Knowledge from the World Wide Web". In: *Proc. of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference, AAAI 98, IAAI 98, July 26-30, 1998, Madison, Wisconsin, USA*. 1998, pp. 509–516.

[9] Michal Rosen-Zvi, and Thomas L. Griffiths, and Mark Steyvers, and Padhraic Smyth. "The Author-Topic Model for Authors and Documents". In: *UAI '04, Proc. of the 20th Conference in Uncertainty in Artificial Intelligence, Banff, Canada, July 7-11, 2004*. 2004, pp. 487–494.

[10] Thomson Reuters. "OpenCalais". In: *Retrieved June* 16 (2008).

[11] Khaled Shaalan. "A Survey of Arabic Named Entity Recognition and Classification". In: *Computational Linguistics* 40.2 (2014), pp. 469–510.