Incorporating Entities in News Topic Modeling

Linmei Hu[1], Juanzi Li[1], Zhihui Li[2], Chao Shao[1], and Zhixing Li[1]

[1] Dept. of Computer Sci. and Tech., Tsinghua University, China
{hulinmei,ljz,shaochao,zhxli}@keg.cs.tsinghua.edu.cn
[2] Dept. of Computer Sci. and Tech., Beijing Information Science and Technology University, China
wisdomlee0606@126.com

Abstract. News articles express information by concentrating on named entities like who, when, and where in news. Whereas, extracting the relationships among entities, words and topics through a large amount of news articles is nontrivial. Topic modeling like Latent Dirichlet Allocation has been applied a lot to mine hidden topics in text analysis, which have achieved considerable performance. However, it cannot explicitly show relationship between words and entities. In this paper, we propose a generative model, Entity-Centered Topic Model(ECTM) to summarize the correlation among entities, words and topics by taking entity topic as a mixture of word topics. Experiments on real news data sets show our model of a lower perplexity and better in clustering of entities than state-of-the-art entity topic model(CorrLDA2). We also present analysis for results of ECTM and further compare it with CorrLDA2.

Keywords: news, named entity, generative entity topic models.

1 Introduction

With the popularization of the Internet, reading online news has become an elementary activity in people's daily life. Named entities which refer to names, locations, time, and organizations play critical roles in conveying news semantics like who, where and what etc. In today's fast-paced life, capturing the semantic relationships between news and entities can help people to understand and explore news. It has abroad applications such as news summarization, multiple dimensional news search and news event extraction.

Recently, there is growing interest and lots of valuable researches in finding and analyzing entities mentioned in news using topic models. In [1], a generative entity-topic model is proposed to combine context information and topics for entity linking given global knowledge. [2] use nonparametric topic models and hierarchial topic models for entity disambiguation respectively. [3] presents a named entity topic model for named entity query, specifically finding related topics given a set of related entities. Though entity topic models have been widely used in many applications, in this paper, we focus on extracting relationships between entities and news article, which can be applied in the applications mentioned above.

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 139–150, 2013.

Figure 1 (a)-(c) illustrates different dependencies between entities and topics of previous topic models. As shown in Figure 1(a), LDA [4] can also be extended to detect interaction between entities and topics, but it cannot model the relationship between words and entities. Figure 1(b) illustrates ETM [5] modeling the generative process of a word given its topic and entity information, but it does not cover relationship among entities. As we can see from Figure 1(c), David Newman et al. propose entity topic model, called CorrLDA2, modeling word topic as a distribution over entity topics [6]. The model can cluster entities and explicitly show word topic related entity topics, while experimental results show unreasonable clustering of entities. In this paper, we propose an entity-centered topic model, ECTM shown in Figure 1(d) to model entity topics as mixtures of word topics. ECTM clusters entities better and obviously mines correlation among entities, words and topic, specifically entities-related word topics. The intuition underlying the idea is when writing a news article, usually person, time and location are determined first, then topics are generated by expanding around these entities.

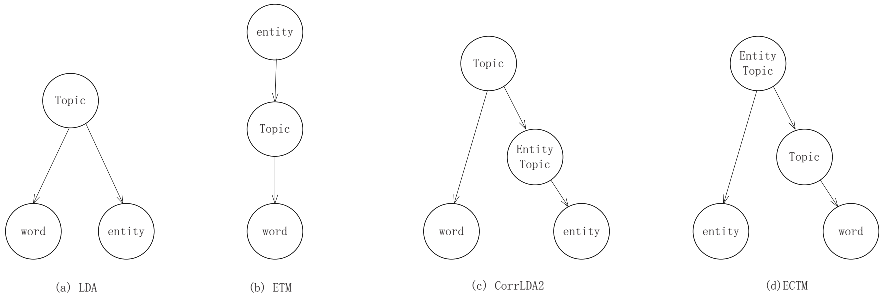


Fig. 1. Different Dependencies in News Modeling

In this paper, with the intuition that entities play a critical role in news articles and the content is usually generated around entities, we propose an entity-centered topic model, ECTM to model entities and topics from a large news articles set. Entity groups and entities-related word topics are demonstrated explicitly.

The main contributions of this paper can be summarized as follows:

- We propose an entity-centered topic model, ECTM to model generation of news articles by sampling entities first, then words.
- We give evidence that ECTM is better in clustering entities than CorrLDA2 by evaluation of average entropy and sKL.
- We apply ECTM on real news articles sets to analyze entity groups, also named entity topics and entities-related word topics. We also present analysis and further compare with entity topic model(CorrLDA2).

The rest of the paper is organized as follows. We review related work in section 2. In section 3, we will define problem and describe ECTM in detail. Experiments and analysis will be given in section 4. Finally, we conclude the paper in section 5.

2 Related Work

In early days, documents are represented as a vector of TF-IDF vlues[7]. In 1990, latent semantic indexing(LSI) [8] was presented to avoid missing related articles but with no keywords occurrence. With further probabilistic theory considered, Hofmann comes up with probabilistic LSI(pLSI)[9] for document modeling. The key idea underlying LDA [4] is that a document is a bag of words, which can be considered as a mixture of latent topics with each topic a distribution over words. Documents are represented as a distribution over K topics, having features reduced compared with TF-IDF representation. In comparison with pLSI, LDA has its merits in parameter estimation and inference.

A lot of topic models have been derived from LDA to meet various scenarios. Author Topic (AT) model [10] incorporates authors in LDA to model authors in different communities, where authors in same community have common interest. In academic searching and analysis, Author Conference Topic (ACT) model is proposed to simultaneously model topical aspects of papers, authors, and publication venues [11]. David Andrzejewski et al. incorporate domain knowledge into LDA to maximize the utilization of domain knowledge [12]. [6] presents a statistical entity topic model, CorrLDA2, modeling entities and topics, where word topics are mixtures of entity topics. However, it cannot meet certain scenarios, for example, presenting entity topic's distribution over word topics directly. Some improvements aim at modeling correlation among topics, such as correlated topic model (CTM) [13], Pachinko allocation model(PAM)[14] and even hierarchical Pachinko allocation model (HPAM)[15], hierarchical LDA (hLDA)[16] for learning hierarchical topic relationships. Gibbs sampling [17] is a widely-used alternative to variational method for topic model's inference.

Named entity has become a catch eye word currently. Various aspects of named entities including named entity recognition, entity linking, entity resolution have been researched. [18] recognizes and finds relationships between named entities in web pages. We address entity-centered summarization of the news articles set by utilizing ICTCLAS [1] for entity recognition. In terms of entity linking, [19] links named entities with knowledge base via semantic knowledge. As our model can enrich semantic information related with the given entity, it definitely can be applied in entity linking.

3 Models

In this section, we describe three graphical entity topic models for news modeling. We begin with existing topic models, LDA and CorrLDA2. Then we discuss

[1] `http://www.nlpir.org/`

about our proposed model, ECTM. All models discussed in this paper treat documents as bag of words. We first introduce some basic notations used in graphical models.

Table 1. Notation Definition

notation	defination
K,$\tilde{K}$	number of word topics, number of entity topics
θ	topic distribution of document
Ψ	super topic distribution over topic distribution
β,$\tilde{\beta}$	word distribution, entity distribution
α,η,$\tilde{\eta}$,μ	prior of θ, β,$\tilde{\beta}$,Ψ
M	number of documents in news articles collection
N_d	number of words and entities in document d
N,$\tilde{N}$	number of words, entities in a document

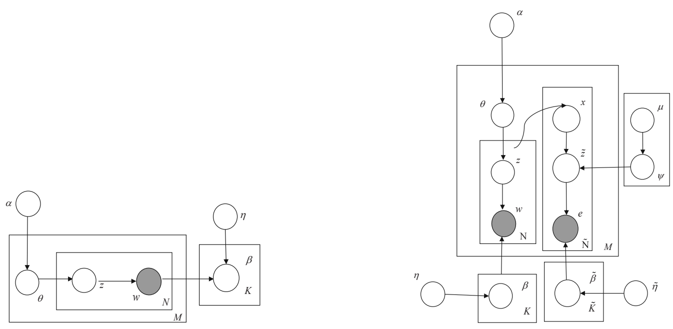
3.1 Previous Models

We illustrate LDA in Figure 2(a) where entities are identical to words. After post-processing, we can extract relationships between entities and topics. LDA models documents mixtures of topics which are mixtures of words. For each word, the model generates a topic according to multinomial distribution. Specifically, it generates a document in the following process:

1. For each $d \in D$
   - sample topic distribution $\theta \sim Dir(\alpha)$
   - for all topics $t = 1, 2, ...K$,sample word distribution $\beta \sim Dir(\eta)$
2. For each word $w \in d$
   - sample a topic $z \sim Multi(\theta)$
   - sample a word $w \sim Multi(\beta)$

Figure 2(b) presents statistical entity topic model, CorrLDA2, where words are sampled before entities. For each entity, it samples a super word topic first, then sample an entity topic. Therefore, the modelling results reveal word topic related different groups of entities. Another difference is that CorrLDA2 allows different topic numbers for word topic and entity topic. The generative process of it is as follows:

1. For all $d \in D$ sample $\theta_d \sim Dir(\alpha)$
2. For all $t = 1, 2, ..., K$ word topics sample $\beta_t \sim Dir(\eta)$ and $\psi_t \sim Dir(\mu)$
3. For all $t = 1, 2, ...\tilde{K}$ entity topics sample $\tilde{\beta}_t \sim Dir(\tilde{\eta})$
4. For each word $w \in d$
   - sample a topic $z \sim Mult(\theta_d)$
   - sample an word $w \sim Mult(\beta_d)$
5. For each entity $e \in d$
   - sample a super word topic $x \sim Mult(z_1, z_2, ..., z_K)$
   - sample an entity topic $\tilde{z} \sim Mult(\psi_x)$
   - sample an entity $e \sim Mult(be\tilde{t}az)$



(a) LDA (b) CorrLDA2

Fig. 2. Graph Representation of Previous Models

3.2 Entity-Centered Topic Model

With the intuition that entity play a pivotal role in news articles, we propose a new entity topic model, entity-centered topic model, ECTM. ECTM is similar with CorrLDA2 but differs from it mainly in sampling order of entity and word. With our model, entity topic contains a mixture of groups of words, for example, an entity topic of earthquake places may include word topics of earthquake disaster situation, personnel recovery, reconstruction of city and tsunami. However, previous models cannot produce such modelling results.

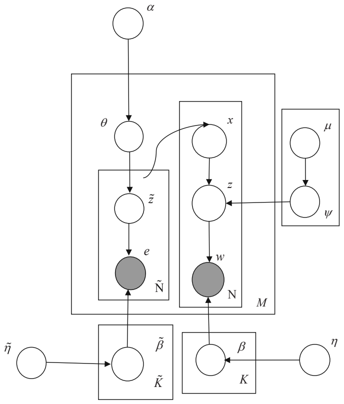


Fig. 3. Graph Representation of ECTM

Figure 3 illustrates our ECTM. In detail, ECTM models a news article following the generative process:

1. For all $d \in D$ sample $\theta_d \sim Dir(\alpha)$
2. For all entity topics $\tilde{t} = 1, 2, ...\tilde{K}$, sample $\tilde{\beta}_t \sim Dir(\tilde{\eta})$ and $\psi \sim Dir(\mu)$
3. For all word topics $t = 1, 2, ...K$, sample $\beta \sim Dir(\eta)$
4. For each entity $e \in d$
   - sample an entity topic $z \sim Multi(\tilde{\theta})$ according to Equation 1
   - sample an entity $e \sim Multi(\tilde{\beta})$
5. For each word $w \in d$
   - sample a super entity topic $x \sim Multi(\tilde{z}_1, \tilde{z}_2, ..., \tilde{z}_{\tilde{K}})$
   - sample a topic $z \sim Multi(\psi)$ according to Equation 2
   - sample a word $w \sim Multi(\beta)$

According to the graphical model of ECTM, we infer the equations of full conditional probabilities used in gibbs sampling as follows.

$$p(\tilde{z}_i = \tilde{k} | e_i = e, \tilde{z}_{\neg i}, e_{\neg i}, \alpha, \tilde{\beta}) \propto \frac{N^{\tilde{k}}_{\tilde{m}_{\neg i}} + \alpha}{\sum_{\tilde{k}'=1}^{\tilde{k}'=\tilde{K}} N^{\tilde{k}'}_{\tilde{m}_{\neg i}} + \tilde{K}\alpha} \cdot \frac{N^{e}_{\tilde{k}_{\neg i}} + \tilde{\beta}}{\sum_{e'=1}^{e'=\tilde{V}} N^{e'}_{\tilde{k}_{\neg i}} + \tilde{V}\tilde{\beta}} . \tag{1}$$

$$p(z_i = k, x = \tilde{k} | w_i = w, z_{\neg i}, \tilde{z}, w_{\neg i}, \beta) \propto \frac{N^{\tilde{k}}_{\tilde{m}} + 1}{N + \tilde{K}} \cdot \frac{N^{k}_{\tilde{k}_{\neg i}} + \mu}{\sum_{k'} N^{k'}_{\tilde{k}_{\neg i}} + K\mu} \cdot \frac{N^{w}_{k_{\neg i}} + \beta}{\sum_{w'=1}^{w'=V} N^{w'}_{k_{\neg i}} + V\beta} . \tag{2}$$

3.3 Time Complexity

Gibbs Sampling is a widely used algorithm for probabilistic models. We assume the iteration times is t and the total words and entities in the document set are N and $\tilde{N}$, then for LDA, the time is $O(t(N + \tilde{N}))$. For CorrLDA2, as for each entity, we will first sample a super word topic, then an entity topic, thus the time is $O(tN + 2t\tilde{N})$. Similarly, ECTM consumes time $O(t\tilde{N} + 2tN)$, where number of words, N is usually much bigger than entity number $\tilde{N}$. Therefore, ECTM is usually more time-consuming than previous models.

4 Experiments

4.1 Experiment Setup

To cover various kinds of news articles set, we collect three data sets from NewsMiner[2]. One is an event about Chile Earthquake called Dataset1, in which all articles are about Chile Earthquake happened in February, 2010. Another is

[2] `http://newminer.net`

about various kinds of intranational news, Dataset2 containing various news in our country. The last data set is a collection about three events including Qinghai Earthquake, Two Sessions in 2013 and Tsinghua University, named Dataset3. Dataset1 and Dataset2 have 632 and 700 news articles respectively, Dataset3 has 1800 articles. We preprocess the data by 1) word segmentation and entity recognition with ICTCLAS which has been mentioned in section 2, 2) removing stop words. Afterwards, there are 5482 words, 1657 entities in Dataset1, 15862 words, 5357 entities in Dataset2 and 19597 words, 10981 entities in Dataset3.

CorrLDA2 has been proven more effective in entity prediction [6]. Therefore, we evaluate ECTM's performance by perplexity taking LDA and CorrLDA2 as baselines [4]. To further analyze the entity topics generated using different models, we measure with average entropy of entity topics computed as Equation 3 and average sKL between each pair of entity topics according to Equation 4. Finally, we analyze and compare overall results of different models.

4.2 Perplexity

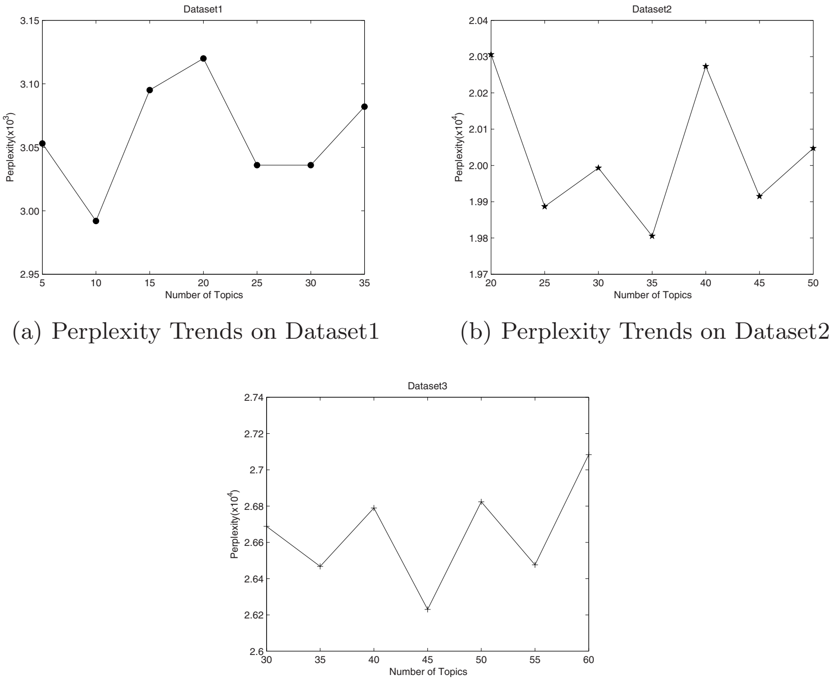
Perplexity is usually used to evaluate a topic model's ability to fit data. All topic models presented in this paper use gibbs sampling for estimation and inference. The prior α is set to be $50/K$, so is μ, $50/\tilde{K}$, β is set to be 0.1, empirically [20]. In terms of topic number K, we empirically determine by choosing the value where the perplexity is lowest. Here, we conduct 10-fold cross validation and take the average perplexity of ten times as final performance. Each time, we choose 90% as training data and the remaining 10% as test data for which we calculate the perplexity. We choose different number of topics to get the perplexity curves of LDA as shown in Figure 4.

We can see from Figure 4, the best topic numbers for three data sets are 10, 35 and 45 respectively. Applying the topic number in CorrLDA2 and ECTM, we determine the number of entity topics where the perplexity is lowest. The best entity topic numbers are 10, 30, 35 for Dataset1, Dataset2, Dataset3 separately. In Table 2, we list corresponding perplexity with three models in different data sets.

Table 2. Perplexity

	LDA	CorrLDA2	ECTM
Dataset1	2991.9535	1508.8008	1412.5654
Dataset2	19805.3647	11593.1167	10415.7223
Dataset3	26230.7089	14891.4707	13544.4122

It can be apparently seen that our proposed model has lowest perplexity, which shows ECTM performs well. The reason behind that may be it is more reasonable for news article writing with entities going first. When one is going to write a news article, he should determine the person, organization, time, location that he's going to write about first, then what happened with the named entities.



(a) Perplexity Trends on Dataset1

(b) Perplexity Trends on Dataset2

(c) Perplexity Trends on Dataset3

Fig. 4. Perplexity of LDA on Topic Trends

4.3 Evaluation of Entity Topic

As both CorrLDA2 and ECTM have entity topic which is a distribution of entities. We set a probabilistic threshold and for each entity topic, we consider entities whose probability is larger than threshold belonging to the entity topic. Then we calculate the average entropy of all entity topics and average sKL for each topic pair to evaluate the entity clustering performance of different models. The results computed according to Equation 3 and Equation 4 are shown in Table 3. The equations of entropy and sKL are shown as below:

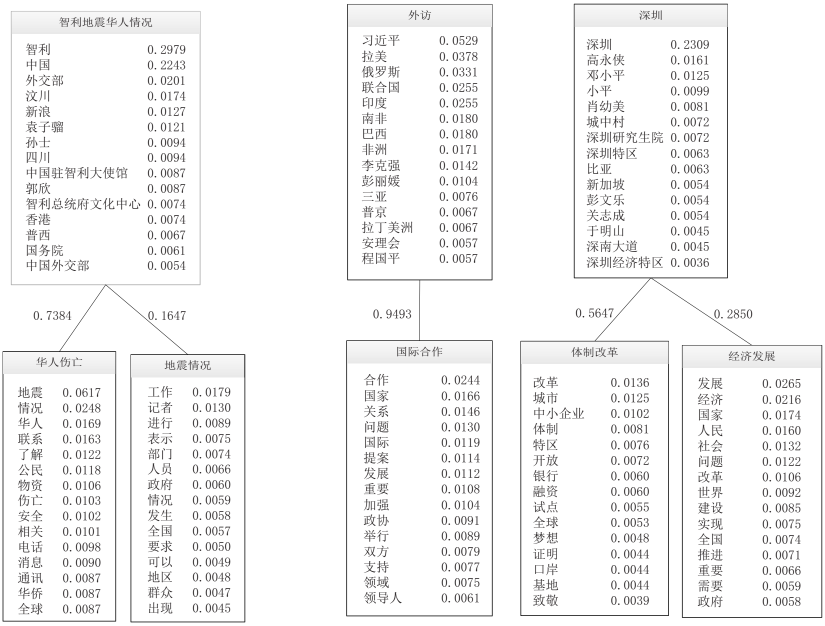
$$Entropyof(Topic) = -\sum_{z} p(z) \sum_{e} p_z(e) \log_2[p_z(e)] \ . \tag{3}$$

$$sKL(i,j) = \sum_{e=1}^{\tilde{N}} [\tilde{\beta}_{ie} \log_2 \frac{\tilde{\beta}_{ie}}{\tilde{\beta}_{je}} + \tilde{\beta}_{je} \log_2 \frac{\tilde{\beta}_{je}}{\tilde{\beta}_{ie}}] \ . \tag{4}$$

As we all know, lower entropy and higher sKL means better clustering of entities with small distance within entity topic and large distance between topics. From Table 3, we find that ECTM performs better than CorrLDA2 except Dataset1 with threshold 0.001 and 0.01 and Dataset 2 with threshold 0.0001. For

Table 3. Entropy and sKL

	Entropy						sKL	
	0.0001		0.001		0.01			
	CorrLDA2	ECTM	CorrLDA2	ECTM	CorrLDA2	ECTM	CorrLDA2	ECTM
Dataset1	0.9540	0.3340	0.0784	0.3244	0.0553	0.2841	7.6710	7.8022
Dataset2	2.4365	3.1221	0.0457	0.0251	0.0710	0.0171	3.2836	4.3278
Dataset3	0.0444	0.0331	0.0427	0.0369	0.0295	0.0219	2.7085	4.2780



(a) Results on Dataset1

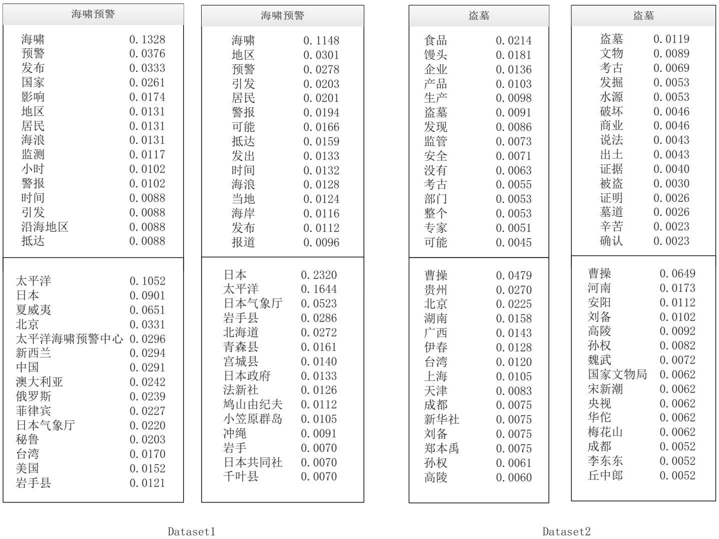
(b) Results on Dataset2

Fig. 5. Results of ECTM

Dataset1, the reason may be entity topics are very similar in an event collection. While for Dataset2, the reason is that the threshold is too low to be reasonable. In terms of sKL, we find for all data sets, even Dataset1, ECTM gets larger sKL between topics.

4.4 Analysis and Comparison of Results

In this section, we will display and analyze the results of topic modeling applied in three different data sets. To show the ability of ECTM to organize news articles as distributions over entity topics, where entity topic is a distribution over word topics, we show part of modeling results on Dataset1 and Dataset2 in Figure 5.



海啸预警		海啸预警	
海啸	0.1328	海啸	0.1148
预警	0.0376	地区	0.0301
发布	0.0333	预警	0.0278
国家	0.0261	引发	0.0203
影响	0.0174	居民	0.0201
地区	0.0131	警报	0.0194
居民	0.0131	可能	0.0166
海浪	0.0131	抵达	0.0159
监测	0.0117	发出	0.0133
小时	0.0102	时间	0.0132
警报	0.0102	海浪	0.0128
时间	0.0088	当地	0.0124
引发	0.0088	海岸	0.0116
沿海地区	0.0088	发布	0.0112
抵达	0.0088	报道	0.0096
太平洋	0.1052	日本	0.2320
日本	0.0901	太平洋	0.1644
夏威夷	0.0651	日本气象厅	0.0523
北京	0.0331	岩手县	0.0286
太平洋海啸预警中心	0.0296	北海道	0.0272
新西兰	0.0294	青森县	0.0161
中国	0.0291	宫城县	0.0140
澳大利亚	0.0242	日本政府	0.0133
俄罗斯	0.0239	法新社	0.0126
菲律宾	0.0227	鸠山由纪夫	0.0112
日本气象厅	0.0220	小笠原群岛	0.0105
秘鲁	0.0203	冲绳	0.0091
台湾	0.0170	岩手	0.0070
美国	0.0152	日本共同社	0.0070
岩手县	0.0121	千叶县	0.0070

Dataset1

盗墓		盗墓	
食品	0.0214	盗墓	0.0119
馒头	0.0181	文物	0.0089
企业	0.0136	考古	0.0069
产品	0.0103	发掘	0.0053
生产	0.0098	水源	0.0053
盗墓	0.0091	破坏	0.0046
发现	0.0086	商业	0.0046
监管	0.0073	说法	0.0043
安全	0.0071	出土	0.0043
没有	0.0063	证据	0.0040
考古	0.0055	被盗	0.0030
部门	0.0053	证明	0.0026
整个	0.0053	墓道	0.0026
专家	0.0051	辛苦	0.0023
可能	0.0045	确认	0.0023
曹操	0.0479	曹操	0.0649
贵州	0.0270	河南	0.0173
北京	0.0225	安阳	0.0112
湖南	0.0158	刘备	0.0102
广西	0.0143	高陵	0.0092
伊春	0.0128	孙权	0.0082
台湾	0.0120	魏武	0.0072
上海	0.0105	国家文物局	0.0062
天津	0.0083	宋新潮	0.0062
成都	0.0075	央视	0.0062
新华社	0.0075	华佗	0.0062
刘备	0.0075	梅花山	0.0062
郑本禹	0.0075	成都	0.0052
孙权	0.0061	李东东	0.0052
高陵	0.0060	丘中郎	0.0052

Dataset2

Fig. 6. Comparison of CorrLDA2 and ECTM

In Figure 5(a), we can see ”智利”, ”中国”, ”外交部” these entities in the entity topic. We can judge the entity topic is about both China and Chile Earthquake. As expected, the entity topic is 73 percent talking about Chinese people situation in earthquake and 16 percent talking about Chile earthquake, which is shown on the edges. The top 15 words in a topic is listed below the topic label. It is unsurprising that we see ”华人” in the word topic under the entity topic with many entities such as ”中国”, ”外交部”, ”中国驻智利使馆” implying China.

In Figure 5(b), ”外访” includes entities like ”习近平”, ”拉美” and is almost only related to the word topic ”国际合作”。However, the entity topic ” 深圳” is almost 60% about ”体制改革” and 30% about ”经济发展” as shown in the figure. From those results, ECTM shows its ability to cluster entities and mine entities-related word topics.

We briefly compare ECTM's experimental results on Dataset1 and Dataset2 with CorrLDA2. The left is result of CorrLDA2 and the right is result of ECTM where result is presented with word topics above of most related group of entities. We choose almost the same word topic and pick up corresponding closely related entity topic representing by top 15 entities. As mentioned previously, ECTM is more powerful than CorrLDA2 in clustering entities and extracting entities-related word topics. We will illustrate this in following Figure 6.

We can see that ECTM clusters entities which are closely related to ”日本”, while corrLDA2 clusters all countries related with ”海啸预警” on Dataset1. For Dataset2, the results of CorrLDA2 are worse. For one side, as ”曹操” should be related with ”河南安阳” where he was buried, the entities under the entity topic are not as relevant as those generated by ECTM. For another side, the words included in the word topic, "盗墓" are also not accurate by the reason of some unrelated words such as "食品", "安全" and so on.

4.5 Experimental Observations

In summary, our model, ECTM has three advantages. Firstly, it has lower perplexity than LDA and CorrLDA2. Secondly, ECTM mines hidden entity topic with better performance, specifically lower entropy and higher sKL than CorrLDA2. Lastly, ECTM extracts entities-related word topics as described in previous subsection, presenting correlation among entities, words and topics. Nevertheless, previous models cannot show entities related words and word topics directly.

5 Conclusions

In this paper, we address the problem of extracting correlation among entities, words and topics. We develop a new entity-centered topic model, ECTM which samples entity first and then samples words according to already-sampled super entity topics. Therefore, ECTM models news collection entity topics which are mixtures of word topics. Through experiments, we find ECTM better in clustering entities and mining entities-related word topics than state-of-the-art entity topic model CorrLDA2.

Entity has become an eye-catching word now. Various aspects about entity has been researched. Incorporating entities into news topic modeling is a challenging problem. There are still many potential future work. For example, we can develop hierarchial entity topic models to mine correlation between entities and topics. We can also research about different models fitting different news collections, specifically, what model fits a news event collection and what model fits collection with various topics.

Acknowledgement. The work is supported NSFC (No. 61035004), NSFC-ANR(No. 61261130588), 863 High Technology Program (2011AA01A207), FP7-288342, THU-NUS NExT Co-Lab, Doctoral Program of Higher Schools(No. 20110002110013), Natural Science Foundation of Inner Mongolia(2011ms0914), Chongqing Science and Technology Research Institute(cstc2011ggB40028).

References

1. Han, X., Sun, L.: An entity-topic model for entity linking. In: EMNLP-CoNLL, pp. 105–115 (2012)
2. Sen, P.: Collective context-aware topic models for entity disambiguation. In: Proceedings of the 21st International Conference on World Wide Web, WWW 2012, pp. 729–738 (2012)
3. Xue, X., Yin, X.: Topic modeling for named entity queries. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM 2011, pp. 2009–2012 (2011)

4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
5. Kim, H., Sun, Y., Hockenmaier, J., Han, J.: Etm: Entity topic models for mining documents associated with entities. In: ICDM 2012, pp. 349–358 (2012)
6. Newman, D., Chemudugunta, C., Smyth, P.: Statistical entity-topic models. In: KDD, pp. 680–686 (2006)
7. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill Book Company (1984)
8. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. JASIS 41(6), 391–407 (1990)
9. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR, pp. 50–57 (1999)
10. Rosen-Zvi, M., Griffiths, T.L., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: UAI, pp. 487–494 (2004)
11. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2008 (2008)
12. Andrzejewski, D., Zhu, X., Craven, M.: Incorporating domain knowledge into topic modeling via dirichlet forest priors. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, pp. 25–32 (2009)
13. Blei, D.M., Lafferty, J.D.: Correlated topic models. In: NIPS (2005)
14. Li, W., McCallum, A.: Pachinko allocation: Dag-structured mixture models of topic correlations. In: Proceedings of the 23rd International Conference on Machine Learning, ICML 2006, pp. 577–584 (2006)
15. Mimno, D.M., Li, W., McCallum, A.: Mixtures of hierarchical topics with pachinko allocation. In: ICML, pp. 633–640 (2007)
16. Blei, D.M., Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B.: Hierarchical topic models and the nested chinese restaurant process. In: NIPS (2003)
17. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(suppl. 1), 5228–5235 (2004)
18. Zhu, J., Uren, V., Motta, E.: ESpotter: Adaptive named entity recognition for web browsing. In: Althoff, K.-D., Dengel, A.R., Bergmann, R., Nick, M., Roth-Berghofer, T.R. (eds.) WM 2005. LNCS (LNAI), vol. 3782, pp. 518–529. Springer, Heidelberg (2005)
19. Shen, W., Wang, J., Luo, P., Wang, M.: Linden: linking named entities with knowledge base via semantic knowledge. In: Proceedings of the 21st International Conference on World Wide Web, WWW 2012, pp. 449–458 (2012)
20. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(suppl. 1), 5228–5235 (2004)