Intelligent Data Analysis 21 (2017) 443–461
DOI 10.3233/IDA-160021
IOS Press

Incorporating Wikipedia concepts and categories as prior knowledge into topic models

Kang Xu, Guilin Qi*, Junheng Huang and Tianxing Wu
School of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu, China

Abstract. Topic models have been widely applied in discovering topics that underly a collection of documents. Incorporating human knowledge can guide conventional topic models to produce topics which are easily interpreted and semantically coherent. Several knowledge-based topic models have been proposed, but these models just leverage lexical knowledge of words that are often not in accordance with topics. To solve the problem, we recognize entity mentions, besides words, in the documents and incorporate entity knowledge from external knowledge bases. In this paper, we study to utilize entity knowledge, concepts and categories in Wikipedia, as prior knowledge into topic models to discover more coherent topics. A novel knowledge-based topic model, WCM-LDA (Wikipedia-Category-concept-Mention Latent Dirichlet Allocation), is proposed, which not only models the relationship between words and topics, but also utilizes concept and category knowledge of entities to model the semantic relation of entities and topics. We compare WCM-LDA with the state-of-the-art knowledge-based topic models, on three datasets. Experimental results show that our approach outperforms the existing baseline methods on all three datasets. Moreover, our model can visualize topics with top words, concepts and categories such that topics are made easily to be interpreted and classified.

Keywords: Knowledge-based topic model, WCM-LDA, Wikipedia thesaurus, entity, concept, category

1. Introduction

Topic models, such as pLSA (Probabilistic latent semantic analysis) [1] and LDA (Latent Dirichlet Allocation) [2], are popular content analysis techniques. They are unsupervised machine learning algorithms to infer latent topics in the corpus and have already had a series of successful applications in topic discovery [3], opinion mining [4], user interest profiling [5], news mining [6,7], graph mining [8].

Topic models take text documents as input, where each document should be split into a bag of words and set the number of topics K. Based on the objective functions of topic models, the models will produce K topics where each topic is represented by a group of words. The objective of topic models is to discover coherent topics in accordance with human judgements, so human judgement is a very important method to evaluate the topics discovered by topic models [9]. But human judgement is time-consuming and labour intensive, we need some automatic evaluation methods. Previously, topic models have been evaluated by computing the perplexity of a hold-out test set. The perplexity is a measurement of how well a probability distribution or probability model predicts a hold-out test set. A low perplexity

*Corresponding author: Guilin Qi, School of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu, China. E-mail: gqi@seu.edu.cn.

indicates the probability distribution is good at predicting the hold-out test set [2]. However, perplexity on the holdout test set does not reflect the semantic coherence of topics [10] and may not be in accordance with human judgements. Therefore, the metric Topic Coherence was proposed in [9] and has shown in [3] to correlate well with human judgements.

Topic models are purely data-driven where topics are generated based on the objective functions. However, some researchers find that the produced topics may not conform to human judgements [9]. One key problem is that the objective functions of topic models (e.g., LDA), which are fully based on implicit word co-occurrence patterns [3], often do not correlate well with human judgements [10].

To solve the above problem, some researchers try to leverage knowledge-based models which incorporate external knowledge to guide topic modeling. The existing knowledge-based topic models mainly incorporate lexical knowledge. For example, DF-LDA (Dirichlet Forest-Latent Dirichlet Allocation) [11] utilizes domain-specific lexical knowledge in the form of must-links and cannot-links to discover topics in accordance with domain knowledge. Two words in a must-link are assigned to the same topic and two words in a cannot-link must be assigned to two different topics. Chen et al. [12–16] further expand domain-specific lexical knowledge to multi-domain knowledge. In their work, words in the same semantic-set share the same word sense. Each semantic-set is viewed as a must-link.

All these models treat each document as "bag of words" and incorporate lexical prior knowledge, but lexical knowledge is often not in accordance with topics of documents. For example, in a document "machine learning is good.", if we only utilize lexical knowledge of words "machine" and "learning", the document may be assigned to the wrong topic "Mechanics" or "Education". However, entities (e.g., a person, a location and a discipline") in the documents are ignored in these topic models. If we recognize the entity mention (a phrase corresponding to an entity) "machine learning" and link the entity mention to a concept [17] (a Wikipedia article that describes an entity), e.g., "Machine learning"@en, we can accurately know that the document is associated with the topic "Computer science". Hence recognizing entity mentions and utilizing the external knowledge of entity mentions can discover coherent and accurate topics. Moreover, if we visualize topics with words, concepts and other knowledge from external knowledge bases, the discovered topics will be easily interpreted by human [18]. For example, if we show a topic with not only words "machine" and "learning", but also a concept "machine learning"@en and a category [19] "Artificial Intelligence"@en, the topic can be more easily interpreted. CorrLDA2 [20], ECTM (Entity-centered Topic Model) [21] and EETM (Event Entity Topic Model) [7] only incorporate some knowledge, such as entity mentions and types of entity mentions (e.g., "Person", "Organization"), to discover more coherent topics. However, these approaches only employ words and entity mentions which can be directly extracted from documents and they don't take advantage of knowledge of entity mentions from external knowledge bases (e.g., concepts in Wikipedia and categories which are classes of concepts).

Entity linking [22] is an intuitive way to link entity mentions to concepts in a given knowledge base. Here we choose Wikipedia, a very dynamic and rapidly growing knowledge base, where a concept corresponds to a Wikipedia article which describes an entity. Besides these concepts, category knowledge in Wikipedia can also be utilized to model topics, because a category is a class of concepts and concepts in the same category have similar semantics. If two documents do not share any words or concepts, but they share two concepts that belong to the same category, then the two documents are more likely to share the same topic. So we utilize these concepts and categories to enrich the documents.

The objective of our work is, by incorporating Wikipedia concepts and categories of entities as prior knowledge, to produce coherent topics. However, there are two key challenges here: *Entity Mention Ambiguity*: For polysemy, an entity mention often corresponds to multiple concepts in Wikipedia among

the multi-domain corpus. For example, an entity mention "Apple" may be linked to different concepts in Wikipedia, such as "Apple Inc." and "Apple Bank for Savings". *Multiple Categories*: When we acquire an unambiguous concept for an entity mention, it also corresponds to multiple categories. However, only a few categories corresponding to the concept could benefit to topic modeling. For example, "Barack Obama" contains categories "Obama family", "1961 births" and "African-American politicians", but in the political topics, only category "African-American politicians" is beneficial for better understanding the topic.

In this paper, we propose a new topic model, called WCM-LDA (Wikipedia-Category-concept-Mention Latent Dirichlet Allocation). There are two kinds of observed variables in our model, entity mentions m that represent entities in the documents and words w. To deal with the problems of entity mention ambiguity and multiple categories in incorporating concepts and categories of entity mentions, we then separately add two latent variables, concept e and category g. Each document has a distribution of latent topics while each topic of words has a probability distribution over words and each topic of entity mentions has a probability distribution over concepts and categories. WCM-LDA is able to handle entity mention ambiguity because the new latent variable e enables the model to choose the right concept from the candidate concepts. Similarly, WCM-LDA can handle multiple categories problem with the new variable g which enables the model to choose the accurate category from the possible categories.

The contributions of this article include:

1) We introduce a general-domain knowledge base, Wikipedia, and utilize the concept and category information as prior knowledge to model topics of multi-domain text documents.
2) We propose a unified topic model which can model topics over words, entity mentions of all the corpus. Thereinto, concepts and categories are incorporated into our model simultaneously to enhance topic modeling. At the same time, our model solves the problems of entity mention ambiguity and multiple categories caused by incorprating concepts and categories.
3) We conduct experiments on three datasets to evaluate the effectiveness of our proposed model and visualize topics with words, concepts and categories.

The rest of the paper is organized as follows: We first discuss related work in Section 2. Then we introduce the background of our work in Section 3. We give the formal definition of our model in Section 4. Then we describe our proposed model WCM-LDA by incorporating Wikipedia concepts and categories as prior knowledge in Section 5. WCM-LDA can model entity mentions, concepts, categories and words simultaneously. In order to evaluate the effectiveness of our model, we conduct experiments in Section 6 and experimental results show our model outperforms the state-of-the-art baselines. Finally we conclude the paper and describe the future work in Section 7.

2. Related work

Topic Models, e.g. pLSA [1] and LDA [2], model semantic relations among words in an unsupervised way. The primitive topic models do not introduce any prior knowledge or other external resources, and topic models produce topics with uncontrolled quality. So some researchers proposed to incorporate external resources, such as social and auxiliary semantics [23], temporal, user and hashtag information [24], and sentiment knowledge [4], into topic models to guide topic generation. But these resources are only available in the constrained domains.

Meanwhile, some other researchers tried to leverage the domain knowledge of words to promote topic modeling. Different from our work that combined entity knowledge, these work all incorporated lexical

(word) knowledge into topic models. The DF-LDA topic model [11] used tree-based priors to encode domain-specific expert knowledge on topic models in the form of must-links and cannot-links. Thereinto, a must-link indicates that two words must be assigned to the same topic, while a cannot-link states that two words should not be in the same topic. In [25], a factor graph framework was proposed to incorporate prior knowledge into topic models, where prior knowledge is modeled as sparse constraints (must links and cannot links) to speed up model training. In [26,27], they all incorporated domain knowledge into topic modeling in the form of first-order logic. Similarly, in [28], Concept Topic Model incorporated domain ontology knowledge to model documents as a mixture of topics and concepts. In [29–33], sets of seed words (i.e., must-links) were introduced as prior knowledge to bias the topic assignment of words. There are some other related work, LDAWN (Latent Dirichlet Allocation with WORDNET) [34] incorporated WordNet-Walk into topic modeling for word sense disambiguation where LDAWN assumes that a word is generated by a WordNet-Walk in WordNet.

Recently, Chen et al. [12–16] proposed a series of research work that incorporated prior lexical knowledge from multi-domains into topic models. Thereinto, the first related work is MDK-LDA (LDA with Multi-Domain Knowledge) [12], a framework that exploited prior knowledge (must-links) from the past domains in topic models to bias topic assignment in the new domains. MC-LDA (LDA with M-set and C-set) [16] is the extension of MDK-LDA which used must-links and cannot-link prior knowledge. In GK-LDA (General Knowledge based LDA) [15], the model not only incorporated the general prior knowledge, but also handled incorrect knowledge without user input. Further they proposed AKL (Automated Knowledge LDA) [13] and LTM (Lifelong Topic Model) [14] that learned prior knowledge automatically from multiple domains to produce more coherent topics. All these works also only incorporated the prior lexical knowledge of words, rather than entities. The prior lexical knowledge often introduce incorrect knowledge which lead to decrease performance of topic modeling.

A list of most probable words is often used to describe individual topics, yet these words often provide a hard-to-interpret or ambiguous representation of the topic. Augmenting words with a list of probable entities provides a more intuitively understandable and accurate description of a topic. So several researchers utilized entities, besides words, to model topics. In [20,21,35], the authors proposed to mine topics for collections of documents explicitly or implicitly associated with sets of entities. The result showed that capturing the association of documents with real-world entities or concepts can enhance the quality of topic modeling. ETM (Entity Topic Model) [35] proposed an entity-topic model for documents associated with entities to model the relationship of words, entities and topics. In [7,20,21], they proposed statistical entity-topic models to model the topics over words and entities which can better understand the topics of documents; only entity mentions and entity types were utilized in the models and did not take advantage of rich semantic knowledge of entities from external semantic resources.

3. Background

3.1. Brief review of LDA

In this section, we review LDA [2] briefly. In LDA, a document is viewed as a multinomial distribution over topics and a topic is a multinomial distribution over words. LDA takes text documents as input, where each document where each document should be split into a bag of words and sets the number of topics K. The output of LDA are document-topic distributions and topic-words distributions. For each topic, we can choose words with the top N probability in the topic-word distribution to visualize the topic. The plate notation of LDA is shown in Fig. 1 and the generative process of LDA is as follows:

Table 1
Notations for LDA

Symbol	Description
$\mathbf{Z}$	All topics
$\mathbf{W}$	All words
d	Index of documents
i	Index of words in the documents
k	Index of topics
$z_{d,i}$	Topic of the i-th word in the d-th document
$w_{d,i}$	i-th word in the d-th document
θ_d	Topic distribution of the d-th document
ϕ	Word distribution of the k-th topic
α	Hyper-parameter of the topic distribution of the d-th document
β	Hyper-parameter of the word distribution of the k-th topic

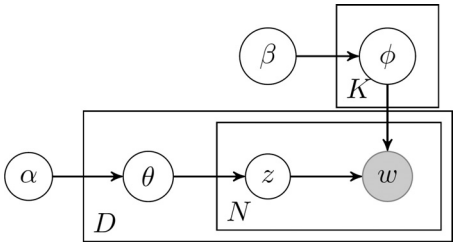

Fig. 1. Plate notation for LDA.

1. For topic k
   (a) Choose a word distribution $\phi_k \sim \text{Dir}(\beta)$
2. For document d
   (a) Choose a topic distribution $\theta_d \sim \text{Dir}(\alpha)$
   (b) For each word w in document d
      i. Choose a topic assignment $z \sim \text{Multi}(\theta_d)$
      ii. Choose a word $w \sim \text{Multi}(\phi_z)$

The total probability of LDA is given in Eq. (1) (The notation are as shown in Table 1):

$$P_{LDA}(\mathbf{Z}, \mathbf{W}, \Theta, \Phi) = \prod_{d,i} p(z_{d,i}|\theta_d) p(w_{d,i}|z_{d,i}, \Phi) \prod_d p(\theta_d) \prod_k p(\phi_k) \tag{1}$$

3.2. Brief introduction of Wikipedia thesaurus

Wikipedia, launched in 2001, is a Web-based, free-content and multilingual encyclopedia, which is edited collaboratively by users around the world. Each article in Wikipedia describes a single concept. Its title is a concise, well-organized phrase that can explicitly represent the article. Each concept must be attached to at least one categories [36]. There exist many hyperlinks between articles, which reflect many semantic relations, such as equivalent relations (synonymy), hierarchical relations (hyponymy) and associative relations [36,37].

Here we introduce the linkage structure in Wikipedia which is a kind of effective knowledge for our model.

Redirect pages: In Wikipedia, each concept is represented by only one article. However, there exist equivalent titles linked to the same concept because they are synonyms. Wikipedia uses a redirect page to link each equivalent title to the source concept. Each redirect page only contains redirect hyperlinks which can handle capitalization, spelling variations, abbreviations, synonyms, colloquialisms and scientific terms.

Disambiguation pages: In Wikipedia, to handle ambiguous terms (e.g., "LDA" may refer to two concepts "Latent Dirichlet Allocation" and "Linear Discriminant Analysis"), disambiguation pages have been created. Disambiguation pages contain various possible meanings. Users can select the intended concepts from disambiguation pages.

Category pages: In Wikipedia, each concept has at least one category. For example, the article "LDA" belongs to two categories "Statistical language processing" and "Latent variable models". Moreover, these categories can be further categorized by associating them with one or more parent categories. In our work, we only leverage the categories that concepts directly link to.

As pointed in [36], Wikipedia also has the following three distinctive advantages, which motivates us to choose Wikipedia as the knowledge base in our work:

(1) It has a very broad knowledge coverage about different concepts from multi-domains, due to the comprehensive contributions by volunteers around the world, which can meet our need of general domain topic discovery;
(2) Its articles are updated regularly, frequently, and consequently, its knowledge repository is always up-to-date;
(3) It contains a large number of new terms that cannot be found in other linguistic corpora, such as WordNet [38], due to its web-based broad participation.

4. Problem definition

In LDA, document generation is interpreted as sampling processing. For each word w in document d, a specific topic z is chosen from the document-topic distribution. Then w is generated according to the topic-word distribution. We can interpret the generative processing of LDA in an intuitive way. When an author wants to write a document, first he chooses a topic he wants to express; then he chooses the most used words that can best express the topic to constitute the content of the document. However, this process only models the words in the documents.

In order to analyze the influence of the entity mentions implied in each document, several variations of LDA have been proposed to model topics over entity mentions and words. These work show that integrating entity mentions into topic models to produce more coherent topics. Because entity mentions reveal extra content implied in documents. Entity mentions can make up this lost semantic information which cannot be completely represented by words. In our work, entity mentions are viewed as a special kind of words in the document and they also contain prior knowledge of the documents from external knowledge bases. Thus, we want to incorporate entity mentions together with concept and category knowledge of these entity mentions to make a modify to the document-topic distribution which eventually produces more coherent topics. In the following, we describe our idea in an intuitive way. Two kinds of tokens (i.e., words and entity mentions) in the documents are generated respectively as follows:

– For each word, the author first chooses a topic he wants to express from the document-topic distribution; then he chooses the words that can represent the topic from the topic-word distribution to generate the content of the document.
– For each entity mention, he also chooses a topic from the document-topic distribution first; then he chooses a category that can describe the topic from the topic-category distribution; and he chooses a concept that belongs to the category and can also best represent the topic from the topic-category-concept distribution; finally he chooses an entity mention that can represent the entity from the entity-mention distribution to generate the content of the document.

Based on the probability distribution of the generative process, we can infer the document-topic distribution, topic-word distribution, topic-concept distribution and topic-category distribution. We choose the words with high probability in the topic-word distribution to represent a topic. Similarly, we can also choose the concepts and categories to represent the topic.

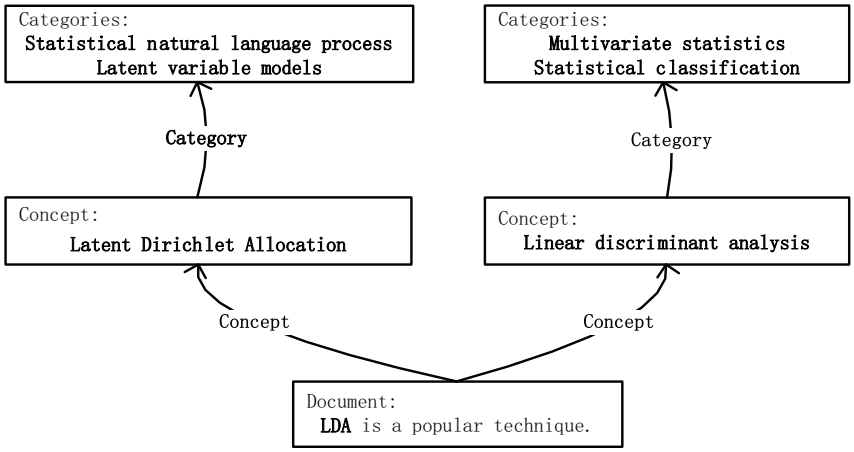

Fig. 2. An example document and prior knowledge for the document.

Here we use an example document to illustrate the generative processing of our model (see Fig. 2). Given a document "LDA is a popular technique", we can recognize an entity mention "LDA" which corresponds to an entity, two words "popular" and "technique". For the entity mention, we have prior knowledge from Concept-Mention Set that "LDA" can be linked to concepts "Latent Dirichlet Allocation" and "Linear Discriminant Analysis". At the same time, we can get category information of these two concepts from Category-Concept Set. Thereinto, "Latent Dirichlet Allocation" corresponds to categories "Statistical language processing", "Latent variable models" and "Linear discriminant analysis" corresponds to categories "Multivariate statistics", "Statistical classification". Based on the content of documents and prior knowledge derived from documents, we can infer topics for each document, and words, concepts, categories for each topic in our work. Before introducing our model, we give some notations used in this paper.

Document. A document contains two kinds of tokens, i.e., words w and entity mentions m. We use d to represent the index of the dth document in the corpus. For example, "LDA is a popular technique." in Fig. 2 contains words "popular" and "technique" and entity mentions 'LDA".

Entity. An entity can be either a person, an organization, a location, or any well-defined notion such as disciplines, nationalities, or wars.

Entity Mention. An entity mention m is a token in document d that represents an entity in the text and m can be a word or a phrase. For example, the word "LDA" in Fig. 2 is an entity mention. An entity mention in different context may correspond to multiple entities.

Concept. A concept e corresponds to a Wikipedia article which describes an entity [17]. For example, "Latent Dirichlet Allocation" in Fig. 2 is a concept. Each entity mention may be linked to multiple concepts for entity mention disambiguation. A concept determinately corresponds to an entity.

Category. A category is a set or class of concepts within a domain [19]. For example, a concept "Latent Dirichlet Allocation" belongs to the category "Latent variable models" in Fig. 2. In Wikipedia, each concept e at least belongs to one category g.

Concept-Mention Set. A Concept-Mention set contains a set of entity mentions linked to the same concept. For example, a concept "Latent Dirchlet Allocation" corresponds to mentions "LDA" and "Latent Dirchlet Allocation".

Category-Concept Set. A Category-Concept set contains a set of concepts that belong to the same category. For example, a category "Latent variable models" contains concepts "Latent Dirchlet Allocation" and "Dynamic topic model".

Mention-Concept-Category Set. A Mention-Concept-Category set contains a set of concept-category pairs describe the same mention. Each mention m has a set of concepts c which can be found in Concept-Mention Sets. At the same time, each concept c has a set of categories which can be found in Category-Concept Sets. So each mention has a set of concept-category pairs. For example, mention "LDA" has a concept-category set: "Latent Dirichlet Allocation" – "Latent variable models" and "Linear Discriminant Analysis" – "Multivariate statistics".

5. Methodology

5.1. Construct mention-concept-category sets for entity mentions

In Wikipedia, all the concept-mention sets and category-concept sets constitute a huge thesaurus. The semantic association of mentions, concepts and categories are explicitly declared in the thesaurus. In Algorithm 1, we introduce the algorithm to construct mention-concept-category sets *EGM* which are prior knowledge of entity mentions. Based on the disambiguation pages and redirect pages, concept-mention sets *EMS* are constructed to look up candidate concepts for each entity mention recognized in the document. For example, both of the terms "Latent Dirichlet Allocation" and "LDA" are all redirected to concept "Latent Dirichlet Allocation", the concept-mention set of the mention "Latent Dirichlet Allocation" is {"Latent Dirichlet Allocation", "LDA"}. Similarly, category-concept sets *GES* are created with category pages to aggregate all the concepts belong to the same category, e.g., the category-concept set of the concept "Latent variable models" {"Latent Dirichlet allocation", "Structural equation modeling"}. For each entity mention in the document, it finds a candidate mention-concept-category set, such as mention "LDA" have the candidate concept-category knowledge set {"Latent Dirichlet Allocation" – "Latent variable models", "Linear Discriminant Analysis" – "Multivariate statistics"}. Detailed steps of Algorithm 1 are explained as follows: for an entity mention m, all the candidate concepts E are found in *EMS*; for each candidate concept e, corresponding categories G are found in *GES* and the candidate concept e is combined with each category g to generate concept-category pairs. All the concept-category pairs are added into a mention-concept-category set EG^m. Hereto, we can acquire concept-category sets *EGM* for all the entity mentions.

Algorithm 1: Construct Mention-Concept-Category Sets for entity mentions

Input: Concept-Mention sets *EMS*, Category-Concept sets *GES*, Entity Mentions m
Output: Mention-Concept-Category Sets *EGM* for all entity mentions m

```
for each entity mention m do
    Initialize a mention-category-concept set EG^m of entity mention m;
    Look up candidate concepts E of mention m in EMS;
    for each candidate concept e in C do
        Look up categories G of concept e in EGS;
        for each category g in G do
            Add concept e and category g pair into EG^m;
        Add EG^m into EGM;
return EGM;
```

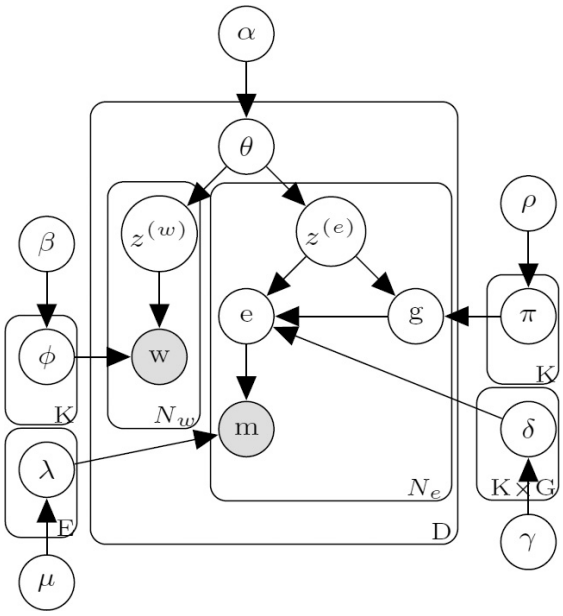

Fig. 3. Plate notation for WCM-LDA.

5.2. Our model

Based on the discussion in Section 3, we propose a model to incorporate Wikipedia thesaurus as prior knowledge into LDA. The plate notation graph (Fig. 3) is given first:

For our model, the full probability is shown in Eq. (2) and the generative process is as follows:

1. For topic k
   (a) Choose a word distribution $\phi_k \sim \text{Dir}(\beta)$
   (b) Choose a category distribution $\pi_k \sim \text{Dir}(\rho)$
   (c) For Category g
      i. Choose a concept distribution $\delta_{kg} \sim \text{Dir}(\gamma)$
2. For concept e
   (a) Choose a mention distribution $\lambda_k \sim \text{Dir}(\mu)$
3. For document d
   (a) Choose a topic distribution $\theta_d \sim \text{Dir}(\alpha)$
   (b) For each word w in document d
      i. Choose a topic assignment $z^{(w)} \sim \text{Multi}(\theta_d)$
      ii. Choose a word $w \sim \text{Multi}(\phi_{z^{(w)}})$
   (c) For each mention m in document d
      i. Choose a topic assignment $z^{(e)} \sim \text{Multi}(\theta_d)$
      ii. Choose a category $g \sim \text{Multi}(\pi_{z^{(e)}})$
      iii. Choose a concept $e \sim \text{Multi}(\delta_{z^{(e)},g})$
      iv. Choose a mention $m \sim \text{Multi}(\lambda_e)$

$$
\begin{aligned}
&P(\mathbf{Z^{(w)}}, \mathbf{Z^{(e)}}, \mathbf{W}, \mathbf{E}, \mathbf{M}, \mathbf{G}) \\
&= \int_{\Theta}\int_{\Phi}\int_{\Delta}\int_{\Lambda}\int_{\Pi} P(\mathbf{Z^{(w)}}, \mathbf{Z^{(e)}}, \mathbf{W}, \mathbf{E}, \mathbf{M}, \mathbf{\Theta}, \mathbf{\Phi}, \mathbf{\Delta}, \mathbf{\Lambda}) d\Pi d\Lambda d\Delta d\Phi d\Theta \\
&= P(\mathbf{Z^{(w)}}, \mathbf{Z^{(e)}}) P(\mathbf{W}|\mathbf{Z^{(w)}}) P(\mathbf{G}|\mathbf{Z^{(e)}}) P(\mathbf{E}|\mathbf{G}, \mathbf{Z^{(e)}}) P(\mathbf{M}|\mathbf{E})
\end{aligned} \tag{2}
$$

In our model, document generation is also interpreted as sampling processing. For each word w in the document d, a specific topic $z^{(w)}$ is chosen from the document-topic distribution θ. Then w is generated according to the topic-word distribution ϕ. For each entity mentions m, a specific topic $z^{(e)}$ is chosen from the document-topic distribution θ, a category g is chosen from the topic-category distribution π, a concept e is chosen from the topic-category-concept distribution δ and finally the mention m is chosen from the concept-mention distribution λ. Thereinto, the concept e and category g of the mention m can be only chosen from the corresponding EG^m. An entity mention is generated by its latent topic, category and concept.

To estimate the parameters for this model, we take the commonly used method, Gibbs sampling [39], to estimate the latent variables. The conditional probabilities of Gibbs sampling in our model are computed by Eqs (3) and (4):

$$
P(Z_i^{(w)} = k|\mathbf{Z}_{-i}^{(\mathbf{w})}, \mathbf{Z^{(e)}}, \mathbf{W_{-i}}, \mathbf{E}, \mathbf{M}, \mathbf{G}) \propto (\alpha + N_{d,-i}^{k,w} + N_d^{k,e}) \frac{\beta + N_{k,-i}^{w}}{W\beta + N_{k,-i}^{*}} \tag{3}
$$

$$
\begin{aligned}
&P(Z_j^{(e)} = k, E_j = e, G_j = g|\mathbf{Z^{(w)}}, \mathbf{Z}_{-j}^{(\mathbf{e})}, \mathbf{W}, \mathbf{E}_{-j}, \mathbf{M}_{-j}, \mathbf{G}_{-j}) \\
&\propto (\alpha + N_d^{k,w} + N_{d,-j}^{k,e}) \frac{\rho + N_{k,-j}^{g}}{G\rho + N_{k,-j}^{*}} \frac{\gamma + N_{k,g,-j}^{e}}{E\gamma + N_{k,g,-j}^{*}} \frac{\mu + N_{e,-j}^{m}}{M\mu + N_{e,-j}^{*}}
\end{aligned} \tag{4}
$$

where "–" indicates excluding that instance from counting. Since we have incorporated prior knowledge from Wikipedia thesaurus, for each mention, the concept e and the category g can be only sampled from the candidate concept-category set of mention m. The notation is as shown in Table 2.

In our work, we want to produce coherent topics and need words, concepts and categories to show the discovered topics. Thereinto, the word distribution (Eq. (6)) and the category distribution (Eq. (7)) over topics can be easily computed. However, in our model, the concept distribution over topics cannot be acquired directly and can be computed in Eq. (9). At the same time, we can get document-topic distribution Eq. (5) and concept-mention distribution Eq. (8).

$$
\theta_{d,k} = \frac{N_d^{k,w} + N_d^{k,e} + \alpha}{N_d^{*} + K\alpha} \tag{5}
$$

$$
\phi_{k,w} = \frac{\beta + N_k^{w}}{W\beta + N_k^{*}} \tag{6}
$$

$$
\pi_{k,g} = \frac{\rho + N_k^{g}}{G\rho + N_k^{*}} \tag{7}
$$

$$
\lambda_{e,m} = \frac{\mu + N_e^{m}}{M\mu + N_e^{*}} \tag{8}
$$

$$
\sigma_{k,e} = \sum_{g=1}^{G} (\pi_k(g).\delta_{k,g}(e)) \tag{9}
$$

Table 2
Notations for our model

Symbol	Description
$\mathbf{Z}, \mathbf{W}, \mathbf{G}, \mathbf{E}, \mathbf{M}$	All topics, words, categories, concepts entity mentions
W, G, E, M	Number of words, categories, concepts entity mentions
$\mathbf{W}_{-i}, \mathbf{E}_{-j}, \mathbf{M}_{-j}, \mathbf{G}_{-j}$	All the words except current word All the categories except current category All the concepts except current concept All entity mentions except current mention
$\alpha, \beta, \rho, \gamma, \mu$	Hyper parameters for our model
$\mathbf{Z}^{(w)}, \mathbf{Z}^{(e)}$	All topic assignment of words, entities
$Z_i^{(w)}, Z_j^{(e)}$	Topic assignment of current words, entities
$\mathbf{Z}_{-i}^{(w)}, \mathbf{Z}_{-j}^{(e)}$	All topic assignment of words except current word All topic assignment of entities except current entity
$N_d^{k,w}, N_d^{k,e}$	Number of topic k over the word w in document d Number of topic k over concept e in document d
$N_{d,-i}^{k,w}, N_{d,-j}^{k,e}$	Number of topic k over the word w in document d except current w Number of topic k over concept e in document d except current concept
$N_{k,-i}^{*}, N_{k,-j}^{*}, N_{k,g,-j}^{*}, N_{e,-j}^{*}$	Number of all words in topic k except current word Number of all categories in topic k except current category Number of all concepts in topic k and category g except current concept Number of all mentions in concept e except current mention m
$N_{k,-i}^{w}, N_{k,-j}^{g}, N_{k,g,-j}^{e}, N_{e,-j}^{m}$	Number of word w in topic k except current word Number of category g in topic k except current category Number of concept e in topic k and category g except current concept Number of mention m in concept e except current mention

5.3. Complexity analysis

Gibbs sampling is a commonly used method for topic models. For time complexity, the iteration times is T, the number of topic is K, the number of the total documents is $|D|$ and the average word length of words for all the documents is $\bar{l}$. The time complexity of LDA is $O(K * T * |D| * \bar{l})$. For our model, the average length of entities for all the documents is $\bar{L}$, the time complexity of our model for all the documents is $O(K * T * |D| * (\bar{l} + \bar{L}))$. Since $\bar{L}$ is also much smaller than $\bar{l}$, the time complexity of LDA and our model are similar. For space complexity, the size of the word vocabulary is W, the space complexity of LDA is $O(|D| * K + K * W + |D| * \bar{l})$. The sizes of concept, category and mention vocabularies are E, G, M, the average size of category for all the topics is $\bar{G}$, the average size of category-concept pairs for all the topics is $\bar{E}$ and the average size of mentions for all the concepts is $\bar{M}$. So the space complexity of our model is $O(|D| * K + K * (W + \bar{G} + \bar{E}) + E * \bar{M} + |D| * (\bar{l} + \bar{L}))$. Since $\bar{E}$, $\bar{M}$ and $\bar{G}$ are all small integers, so the space complexity of our model is not much higher than that of LDA.

6. Experiments

In this section, we evaluated the proposed WCM-LDA model and compared it with five state-of-the-art baseline models:

- LDA [2]: The original generative topic model.
- DF-LDA [11]: A topic model that can introduce human-provided must-link and cannot-link knowledge.

- MDK-LDA [12]: A topic model that can use prior knowledge from multiple domains.
- GK-LDA [15]: A topic model, which is based on MDK-LDA, that can use the ration of words under each topic to reduce the effect of wrong knowledge.
- LTM [14]: A lifelong learning topic model that learn the must-link type of knowledge automatically.

6.1. Datasets and settings

6.1.1. Datasets

Since Wikipedia thesaurus is a general domain knowledge base and the proposed model is domain independent, we used three datasets from different domains to evaluate our model and baseline methods. Reuters-21578 collection[1] contains 12,902 documents which are split into 90 classes, e.g., Earn, Acquisition, Money-fx, etc. In Ohsumed collection,[2] it includes 13,929 medical abstracts from the MeSH categories of the year 1991. These abstracts are classified as the 23 cardiovascular diseases categories. In 20Newsgroups,[3] it contains 19,997 articles for 20 categories (e.g., misc forsale, social religion) taken from the Usenet newsgroups collection.

6.1.2. Pre-processing

First, we used babelfy[4] to recognize the entity mentions and ran the Stanford Part-Of-Speech Tagger[5] to obtain the nouns, verbs and adjectives (for words with other parts-of-speech almost do not indicate topics) for each document in the corpus. Then words appearing less than 5 times in each corpus were also removed. Thereinto, we manually removed some abnormally frequent words and entity mentions (such as "Subject" in the 20Newsgroups) in the corpus. For these words or entity mentions co-occur with most words in the corpus, leading to high similarity among topics. Finally all the left words were converted to lower case.

6.1.3. Wikipedia thesaurus

In our experiments, we used the Jan. 1, 2014 English version of Wikipedia as the prior knowledge, which contains nearly 4.45 million concepts (e.g., "Linear discriminant analysis" and "Latent Dirichlet allocation") and 0.78 millions categories. We used knowledge from the redirection pages and disambiguation pages to construct a variation set for each concept (e.g., the variation set of the concept "Obama" may contain "Barack Obama" and "Barack Hussein Obama II"). Meanwhile, a category set for each concept was constructed from the categories in each Wikipedia pages. Based on the variation sets and category sets of all the concepts, each entity mention recognized before was linked to candidate concepts and categories.

6.1.4. Parameter setting

All models were trained using 1,000 iterations with an initial burn-in of 100 iterations. For all models, the symmetric priors were set as $\alpha = 1$, $\beta = 0.1$ and the numbers of topic were empirically set $K = 20$ (20Newsgroups), 23 (Ohsumed), 90 (Reuters-21578). In DF-LDA, GK-LDA and MDK-LDA, they need prior knowledge from external lexical knowledge base and chose WordNet as the external knowledge

[1] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.
[2] ftp://medir.ohsu.edu/pub/ohsumed.
[3] http://qwone.com/ jason/20Newsgroups/.
[4] http://babelfy.org/.
[5] http://nlp.stanford.edu/software/tagger.shtml.

Table 3
Priors of WCM-LDA

Prior	Value
α	1
β	0.1
π	0.1
μ	0.1
ρ	0.1

Table 4
Topic Coherence across all corpuses and all models

Corpus \ Model	LDA	DF-LDA	MDK-LDA	GK-LDA	LTM	WCM-LDA
20News	−1077.38	−1093.34	−1086.52	−1046.97	−1040.16	−1033.94
Ohsumed	−1052.69	−1065.57	−1035.22	−1040.35	−1031.51	−1011.5
Reuters	−1144.71	−1139.37	−1129.2	−1137.09	−1156.55	−1124.45
Average	−1091.59	−1099.43	−1083.65	−1074.8	−1076.07	−1056.63

Table 5
Cohen's Kappa for pairwise inter-rater agreements

	Topic labeling	Word labeling			
		p@5	p@10	p@15	p@20
Kappa	0.883	0.937	0.891	0.856	0.843

base. So we also chose WordNet for the three models. For DF-LDA, we followed the definition of must-link to generate must-links from WordNet. However, WordNet does not contain cannot-link knowledge, so cannot-link knowledge were ignored here. In WCM-LDA, the symmetric priors π, μ and ρ were empirically set as 0.1 (All the priors of our model are listed in Table 3).

6.2. Topic coherence

Topic models have been evaluated using perplexity traditionally. However, perplexity on the holdout set does not reflect the semantic coherent of topics and is often contrary to human judgments [10]. As our goal is to discover coherent topics, we consider the metric topic coherent correlates well with human judgements [9] which is shown in Eq. (10). So topic coherence is suitable for our evaluation. A higher topic coherence value indicates a higher quality of topics. Thereinto, $D(v)$ is the document frequency of word v, $D(v, v')$ is the co-document frequency of word v and v' and $V^{(t)} = (v_1^{(t)}, \ldots, v_M^{(t)})$ is a list of the M most probable words in topic t. The key idea of the coherence score is that if a word pair is related to the same topic, they will co-occur frequently in the corpus. We followed [9] to calculate topic coherence. Table 4 shows the average topic coherence over all topics in each corpus. Based on the experimental results, our model performed better than baseline methods consistently.

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{10}$$

6.3. Human evaluation

Since we want to discover more coherent topics, we chose to evaluate the topics manually which is based on human judgement. Without enough knowledge, the annotation will not be credible. Following [9], we asked two human judges, who are familiar with common knowledge and skilled in looking up domain knowledge, to annotate the discovered topics manually. To ensure the annotation reliable, we labeled the generated topics by all the baseline models and our proposed model at learning iteration 10.

Topic Labeling: Following [9], we asked the judges to label each topic as *coherent* or *incoherent*. Each topic is represented as a list of 20 most probable words in word distribution ϕ of the topic. Here they annotated a topic as *coherent* when at least half of top 20 words were related to the same semantic-coherent concept (e.g., an event, a hot topic and a discipline), others were *incoherent*. Table 5 shows the

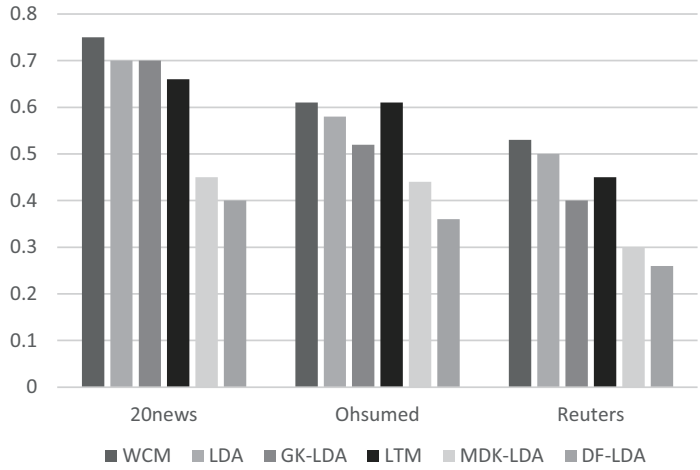

Fig. 4. Proportion of *coherent* topics generated by each model.

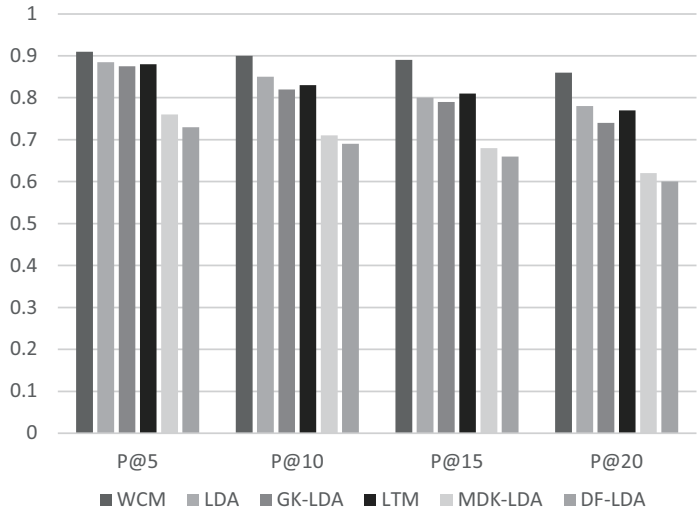

Fig. 5. Average Precision @n (p @n) of words in *coherent* topics generated by each model over three corpuses.

Cohen's Kappa score[6] for topic labeling. When Cohen's Kappa score is greater than 0.8, the annotation of both judges achieve good agreement [40].

Word Labeling: Then we chose *coherent* topics which were judged before and asked judges to label each word of the top 20 words among these *coherent* topics. When a word was in accordance with the main semantic-coherent concept that represents the topic, the word was annotated as *correct* and others were *incorrect.* After topic labeling, the judges had known the concept of each topic, it is easy to label words of each topic. As is shown in Table 5, the annotation of both judges in *Precision*@n (or $p@n$), ($n = 5, 10, 15, 20$), also have good agreements (*Kappa* > 0.8).

Figure 4 shows that WCM-LDA can discover more *coherent* topics than the baseline models. On

[6]https://en.Wikipedia.org/wiki/Cohen%27s_kappa.

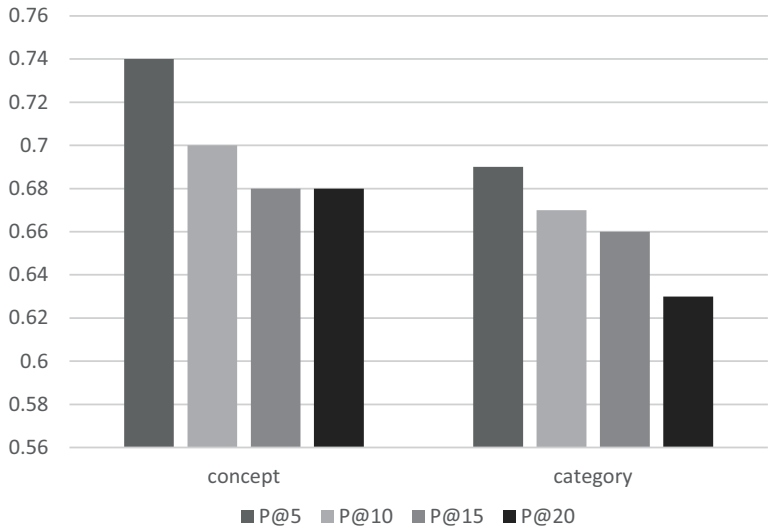

Fig. 6. Average Precision @n (p @n) of concepts and categories in *coherent* topics generated by our model over three corpuses.

average, WCM-LDA discovered about 5% more *coherent* topics than LDA and LTM which are the best baseline models in all the three corpuses.

Figure 5 gives the average *Precision*@n of all *coherent* topics over three corpuses. WCM-LDA outperforms the baseline models and it improves LDA by more than 8% on average. The model makes the most of the high-coverage knowledge of entities from Wikipedia and identified more accurate topics. Thereinto, MDK-LDA and DF-LDA performed worst and even were worse than LDA, the reason we speculate is that, on the one hand, the prior lexical knowledge from WordNet just covers a limited part of real-world knowledge and simultaneously contains some incorrect knowledge. GK-LDA handled the problem of incorrect knowledge, but GK-LDA, like MDK-LDA and DF-LDA, still had the low-coverage problem of prior knowledge. Although LTM alleviated the problem of low-coverage lexical knowledge by automatically learning prior knowledge, it only learned the lexical-level prior knowledge which can not better interpret topics, relative to entity-level knowledge.

In our model, they not only labeled words of *coherent* topics, but also annotated concepts and categories of *coherent* topics referred to word labeling. Average Precision @n (p @n) of concepts and categories in *coherent* topics is shown in Fig. 6. Both precisions of concept and category are more than 0.6 in p@5, p@10, p@15, p@20 and the precision of concept is, on average, 5% more than category. The reason we estimate is that the categories of concepts are collaboratively edited by common users all around the world who mostly have limited range of knowledge and fail to know all the categories of concepts, so a part of concepts lack appropriate categories in accordance with topics discovered from the corpus.

6.4. Example topics

This section shows three example of *coherent* topics produced by WCM-LDA or LDA from three corpuses in an intuitive way. For the convenience of visualization, we showed each topic with the top 10 words here and marked error words in bold. Topics “Christianism”, “Virus Disease” and “Economic and Trade” are shown in Table 6. From the bold words of each topics, we can observe that our model can discover more meaningful topical words than baseline methods. Thereinto, topic “Economic and

Table 6
Three example topics in three corpuses

20news Christianism		Ohsumed Virus Disease		Reuters Economic and Trade	
LDA	WCM-LDA	LDA	WCM-LDA	LDA	WCM-LDA
God	God	*Patients*	Infection	Trade	Market
Jesus	Jesus	Infection	Virus	ec	Markets
Bible	Faith	Virus	Aids	U.S	Traders
Church	Christian	*Human*	HIV	*Community*	Dealers
Christian	Bible	HIV	*Human*	*European*	Trading
People	*Life*	Aids	Immuno-deficiency	Tax	Buying
Love	Church	Antibodies	Hepatitis	Gatt	Expected
Life	*Book*	Immuno-deficiency	Infected	Ministers	Dealer
Paul	Christians	Antibody	Antibody	*Countries*	Demand
Sin	Christ	Serum	HIV-1	*Yeutter*	Analysts

Table 7
Topic "Christianism" in 20news with top 10 words, concepts and categories

20news Christianism		
Word	Concept	Category
God	Authorship of the bible	Christian terms
Jesus	Divinity	Christian biblical canon
Faith	Keith (given name)	Religious belief and doctrine
Christian	Opinion	Scottish masculine given names
Bible	World	Critical thinking
Life	Faith	Epistemology
Church	Life history (sociology)	Concepts in metaphysics
Book	Church fathers	Belief
Christians	Truth	Human development
Christ	Church service	Living people

Table 8
Topic "Virus Disease" in ohsumed with top 10 words, concepts and categories

Ohsumed Virus Disease		
Word	Concept	Category
Infection	Transmission (medicine)	Epidemiology
Virus	HIV-AIDS	Virology
Aids	Virus	Syndromes
Hiv	HIV	HIV-AIDS
Human	Immunodeficiency	Pathology
Immunodeficiency	Molar concentration	Immunodeficiency
Hepatitis	Antibody	Sexually transmitted diseases and infections
Infected	Host (biology)	Diseases and disorders
Antibody	Hepatitis	Chemical properties
Hiv-1	Subtypes of HIV	Viral diseases

Trade" was discovered only by our proposed WCM-LDA, but not LDA. So we tried our best to find the most similar topic from LDA. In our model, we can get not only the words to show the discovered topics, but also the concepts and categories of topics. It is difficult for us to interpret the topics solely with words of each topic, but the concepts and categories are explicit semantic units which hold explicit semantic descriptions in Wikipedia. Concepts and categories of each topic make users easier to interpret

Table 9
Topic "Economic and Trade" in Reuters with top 10 words, concepts and categories

Reuters Economic and Trade		
Word	Concept	Category
Market	Trade	Trade
Markets	Import	International trade
Traders	Economic surplus	Microeconomics
Dealers	South Korea	Member states of the United Nations
Trading	Export	Business terms
Buying	Economic shortage	Commercial item transport and distribution
Expected	Commodity	Marketing
Dealer	Duty (economics)	International economics
Demand	Visible balance	International law
Analysts	General agreement on tariffs and trade	Commerce

topics. As is shown in Tables 7–9, we visualized the aforementioned three topics with topical words, concepts and categories. For example, in Table 7, we can easily interpret the topic "Christianism" from the concept "Authorship of the Bible" and categories "Christian terms", "Religious belief and doctrine". Similarly, topic "Virus Disease" contains concept "Transmission (medicine)" and category "Epidemiology" (Table 8), topic "Economic and Trade" contains concept "Trade" and category "Trade" (Table 9). In summary, our model can not only improve the quality of topics but also make the discovered topics more interpretable. These information can also be utilized for further topic classification.

7. Conclusion and future work

This paper proposed a novel framework to exploit prior entity knowledge of general domain from Wikipedia thesaurus for producing better topics. The existing knowledge-based topic models mainly incorporated lexical knowledge, we innovatively leveraged entity knowledge from Wikipedia thesaurus. To perform the task, the paper identified two key challenges, entity mention ambiguity and multiple categories. A novel model called WCM-LDA was proposed to handle the problems. Experimental results showed the effectiveness of our proposed WCM-LDA. Another advantage of the work is the evaluation on three general-domain corpuses which can reflects the generality of our model. In our model, we visualized topics with top words, concepts and categories. Thus, topics are made more easily to be interpreted and classified.

In our work, we only leveraged concept and category knowledge of entities to guide topic modeling. Other knowledge in knowledge base, e.g., property and disjointness relation [41], are not utilized to guide topic modeling. In the next step, we consider to incorporate these knowledge structure into our model and analyze which knowledge of entities is more effective for topic modeling.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 61672153 and the 863 Program under Grant No. 2015AA015406.

References

[1] T. Hofmann, Probabilistic latent semantic indexing, in: *Proceedings of the 22nd Annual International ACM SIGIR Con-*

ference on Research and Development in Information Retrieval, ACM (1999), 50–57.
[2] D.M. Blei, A.Y. Ng and M.I. Jordan, Latent dirichlet allocation, *J Mach Learn Res* **3** (2003), 993–1022.
[3] X. Yan, J. Guo, Y. Lan and X. Cheng, A biterm topic model for short texts, in: *Proceedings of the 22nd International Conference on World Wide Web*, Springer (2013), 1445–1456.
[4] C. Lin and Y. He, Joint sentiment/topic model for sentiment analysis, in: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ACM (2009), 375–384.
[5] J. Weng, E.-P. Lim, J. Jiang and Q. He, Twitterrank: Finding topic-sensitive influential twitterers, in: *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, ACM (2010), 261–270.
[6] L. Hou, J. Li, Z. Wang, J. Tang, P. Zhang, R. Yang and Q. Zheng, Newsminer: Multifaceted news analysis for event search, *Knowl Based Syst* **76** (2015), 17–29.
[7] L. Hu, C. Shao, J. Li and H. Ji, Incremental learning from news events, *Knowl Based Syst* **89** (2015), 618–626.
[8] J. Xuan, J. Lu, G. Zhang and X. Luo, Topic model for graph mining, *IEEE T Cy* **45** (2015), 2792–2803.
[9] D. Mimno, H.M. Wallach, E. Talley, M. Leenders and A. McCallum, Optimizing semantic coherence in topic models, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ACL (2011), 262–272.
[10] J. Chang, S. Gerrish, C. Wang, J.L. Boyd-graber and D.M. Blei, Reading tea leaves: How humans interpret topic models, in: *Advances in Neural Information Processing Systems*, MIT Press (2009), 288–296.
[11] 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*, ACM (2009), 25–32.
[12] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos and R. Ghosh, Leveraging multi-domain prior knowledge in topic models, in: *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, AAAI (2013), 2071–2077.
[13] Z. Chen, A. Mukherjee and B. Liu, Aspect extraction with automated prior knowledge learning, in: *Proceedings of the 52rd Annual Meeting of the Association for Computational Linguistics*, ACL (2014), 347–358.
[14] Z. Chen and B. Liu, Topic modeling using topics from many domains, lifelong learning and big data, in: *Proceedings of the 31st International Conference on Machine Learning*, ACM (2014), 703–711.
[15] 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 Conference on Information & Knowledge Management*, ACM (2013), 209–218.
[16] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos and R. Ghosh, Exploiting domain knowledge in aspect extraction., in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ACL (2013), 1655–1667.
[17] X. Hu, X. Zhang, C. Lu, E.K. Park and X. Zhou, Exploiting Wikipedia as external knowledge for document clustering, in: *Proc of KDD*, ACM (2009), 389–396.
[18] A. El-Kishky, Y. Song, C. Wang, C.R. Voss and J. Han, Scalable topical phrase mining from text corpora, *PVLDB* **8**(3) (2014), 305–316.
[19] T. Wu, S. Ling, G. Qi and H. Wang, Mining type information from chinese online encyclopedias, in: *Semantic Technology – 4th Joint International Conference, JIST 2014, Chiang Mai, Thailand, November 9–11, 2014*, Revised Selected Papers, Springer (2014), 213–229.
[20] D. Newman, C. Chemudugunta and P. Smyth, Statistical entity-topic models, in: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM (2006), 680–686.
[21] L. Hu, J. Li, Z. Li, C. Shao and Z. Li, Incorporating entities in news topic modeling, in: *Proceedings of the 4th Conference of Natural Language Processing and Chinese Computing*, Springer (2013), 139–150.
[22] X. Han and L. Sun, An entity-topic model for entity linking, in: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ACL (2012), 105–115.
[23] J. Vosecky, D. Jiang, K.W.-T. Leung, K. Xing and W. Ng, Integrating social and auxiliary semantics for multifaceted topic modeling in twitter, *ACM Trans Internet Technol* **14** (2014), 27–50.
[24] Q. Zhang, Y. Gong, X. Sun and X. Huang, Time-aware personalized hashtag recommendation on social media, in: *Proceedings of the 25th International Conference on Computational Linguistics*, ACM (2014), 203–212.
[25] Y. Yang, D. Downey, I. Evanston, J. Boyd-Graber and J.B. Graber, Efficient methods for incorporating knowledge into topic models, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ACL (2015), 308–317.
[26] 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 21st International Joint Conference on Artificial Intelligence*, AAAI (2011), 1171–1192.
[27] S. Mei, J. Zhu and J. Zhu, Robust regbayes: Selectively incorporating first-order logic domain knowledge into bayesian models, in: *Proceedings of the 31st International Conference on Machine Learning*, ACM (2014), 253–261.
[28] C. Chemudugunta, A. Holloway, P. Smyth and M. Steyvers, Modeling documents by combining semantic concepts with unsupervised statistical learning, in: *Proceedings of the 7th International Conference on the Semantic Web*, IEEE (2008), 229–244.

[29] J. Jagarlamudi, H. Daumé III and R. Udupa, Incorporating lexical priors into topic models, in: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, ACL (2012), 204–213.
[30] N. Burns, Y. Bi, H. Wang and T. Anderson, Extended twofold-lda model for two aspects in one sentence, in: *Proceedings of the 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Springer (2012), 265–275.
[31] B. Lu, M. Ott, C. Cardie and B.K. Tsou, Multi-aspect sentiment analysis with topic models, in: *Proceedings of the 11th IEEE International Conference on Data Mining Workshops*, IEEE (2011), 81–88.
[32] A. Mukherjee and B. Liu, Aspect extraction through semi-supervised modeling, in: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL (2012), 339–348.
[33] D. Andrzejewski and X. Zhu, Latent dirichlet allocation with topic-in-set knowledge, in: *Proceedings of the NAACL-HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, NAACL (2009), 43–48.
[34] J.L. Boyd-Graber, D.M. Blei and X. Zhu, A topic model for word sense disambiguation., in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ACL (2007), 1024–1033.
[35] H. Kim, Y. Sun, J. Hockenmaier and J. Han, Etm: Entity topic models for mining documents associated with entities, in: *Proceedings of the IEEE 12th International Conference on Data Mining*, IEEE (2012), 349–358.
[36] G. Xu, Z. Wu, G. Li and E. Chen, Improving contextual advertising matching by using Wikipedia thesaurus knowledge, *Knowl Inf Syst* **43** (2014), 599–631.
[37] D. Milne, O. Medelyan and I.H. Witten, Mining domain-specific thesauri from Wikipedia: A case study, in: *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, IEEE (2006), 442–448.
[38] G.A. Miller, Wordnet: a lexical database for english, *Commun ACM* **38**(11) (1995), 39–41.
[39] G. Casella and E.I. George, Explaining the gibbs sampler, *Amer Statist* **46**(3) (1992), 167–174.
[40] J. Landis and G. Koch, The measurement of observer agreement for categorical data, *Biometrics* **33** (1977), 159–174.
[41] Y. Ma, H. Gao, T. Wu and G. Qi, Learning disjointness axioms with association rule mining and its application to inconsistency detection of linked data, in: *Proceedings of the 8th Chinese Conference of Semantic Web and Web Science*, Springer (2014), 29–41.