Pattern Recognition Letters 123 (2019) 53–60

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Topic representation: Finding more representative words in topic models[☆]

Jinjin Chi[a], Jihong Ouyang[a], Changchun Li[a], Xueyang Dong[c], Ximing Li[a,*], Xinhua Wang[a,b]

[a] *College of Computer Science and Technology, Jilin University, Changchun 130012, China*
[b] *Changchun Institute of Optics Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130012, China*
[c] *Public Computer Education and Research Center, Jilin University, Changchun 130012, China*

ARTICLE INFO

Article history:
Available online 18 March 2019

MSC:
41A05
41A10
65D05
65D17

Keywords:
Topic modeling
Topic representation
Topical word representation
Reranking methodology

ABSTRACT

The top word list, i.e., the top-M words with highest marginal probabilities in a given topic, is the standard topic representation in topic models. Most of recent automatical topic labeling algorithms and popular topic quality metrics are based on it. However, we find, empirically, words in this type of top word list are not always representative. The objective of this paper is to find more representative top word lists for topics. To achieve this, we rerank the words in a given topic by further considering marginal probabilities on words over every other topic. The reranking list of top-M words is used to be a novel topic representation for topic models. We investigate three reranking methodologies, using (1) standard deviation weight, (2) standard deviation weight with topic size and (3) Chi Square χ^2 statistic selection. Experimental results on real-world collections indicate that our representations can extract more representative words for topics, agreeing with human judgements.

1. Introduction

Probabilistic topic modeling family [1] has become a mainstream tool for analyzing the text document collection. This model family assumes that each document is a mixture of latent topics, where each topic is a multinomial distribution over the vocabulary.

Topic models such as latent Dirichlet allocation [2] have empirically achieved great success in modeling documents so far. With algorithms for approximating posterior inference, we can use topic models to uncover latent variables with respect to topics from a collection of documents, leading to semantically meaningful decompositions of them. Topics place high probabilities on words to represent concepts, and documents are described by mixtures of these concepts. Due to the success in discovering semantics knowledge, topic models are usually used in natural language processing tasks, such as multi-document summarization [3] and novel word sense detection [4].

Perusing the learnt topics is the core mission in topic modeling for documents. The standard topic representation is the top word list, i.e., the top-M words with highest marginal probability in a given topic. To simplify the topic representation, some attempts aim at automatically labeling topics, generating labels that can explicitly identify the semantics of topics. For example, the first attempt, to our knowledge, is proposed in [5], which generates candidate labels from a reference collection using noun chunks and bigrams with high lexical association. The authors of [6] suggest an automatical topic labeling algorithm using Wikipedia article titles to process candidate labels. More recently, an interesting algorithm [7] represents topics by image labels, instead of text labels. PageRank is used to select the most suitable candidate images.

Among these existing topic representations, the top word list is acknowledged to be the basic representation for topics, and most of the alternatives, i.e., word cloud [8] and t-SNE [9], to our knowledge, are based on it. More broadly, the automatical topic quality evaluation metrics [10–12] are also based on the top word list. For example, the popular topic coherence metric is computed by counting the co-occurrence numbers among top words in a given topic, following the intuition that more frequently co-occurring intends more coherent a topic is.

All algorithms based on the top word list mentioned above, i.e., both alternative topic representations and topic quality metrics, follow a basic assumption that the top-M words ranked by the topic-word distributions are the most representative words for topics. However, that is not always the case [13]. With statistics inference algorithms, the words with highest marginal probabilities should be high frequency words, no matter whether these words are on a specific subject. At worst, a very frequently occurring but

[☆] **Conflict of interest:** None.
[*] Corresponding author.
E-mail address: liximing86@gmail.com (X. Li).

https://doi.org/10.1016/j.patrec.2019.01.018
0167-8655/© 2019 Elsevier B.V. All rights reserved.

Table 1
Two top-10 word lists of corresponding topics learnt by LDA and MGCTM, respectively. The first and third rows are topics learnt by LDA; the second and fourth rows are topics learnt by MGCTM. The first column shows topic coherence values.

Topic coherence	Top-10 word list
−255.7	good, night, day, morning, sleep, time, work, $\boxed{\text{lol}}$, today, home
−256.5	good, day, morning, night, today, time, work, tomorrow, sleep, home
−282.6	iphone, ipad, apple, free, app, $\boxed{\text{rt}}$, android, phone, online, buy
−297.4	iphone, ipad, apple, phone, follow, free, app, android, buy, online

meaningless word will be the top word for most of topics, just like the stopwords. Some examples can be seen in the next section. Unfortunately, due to the power-law characteristics of language, it is impossible to make a clean sweep of such stopword-like words in practice.

To address the problem mentioned above, we aim to find more representative words for topics, and use the novel list of top-M words to represent topics. This objective can be achieved by reranking the words in a given topic by further considering marginal probability on words over every other topic. Considering two cases: (1) if a top word in a given topic is also in the top word lists for most of the other topics, this word is a bad top word, which is in fact unrepresentative; (2) if a word is not occurring very frequently but most of the occurrences are assigned to the same topic, this word might be representative for that topic. Following the analysis, we propose three reranking methodologies, using (1) standard deviation weight, (2) standard deviation weight with topic size and (3) Chi Square χ^2 statistic selection. We conduct a number of experiments to evaluate our algorithms on real-world collections. Empirical results indicate that in contrast to the standard top word list with highest marginal probability, our algorithms can extract more representative words for topics, agreeing with human judgements.

2. Problem description

In this section, we present the problems in the standard top word list representation. We first review some preliminaries, including two topic models and a popular topic quality metric. The studied topic models are latent Dirichlet allocation (LDA) [2] and Multi-grain clustering topic model (MGCTM) [14]; and the topic quality metric is the so called topic coherence [11,15].

2.1. Preliminaries

LDA. LDA is a generative probabilistic model for the text document collections. This model consists of K topics, where each topic is a multinomial distribution ϕ over the vocabulary, drawn from a Dirichlet prior β. To generate a document d, LDA first draws a topic mixture proportion θ_d from a Dirichlet prior α, and then repeatedly generates word tokens by sampling a topic indicator z_{dn} from the distribution θ_d and then sampling a word w_{dn} from the selected topic distribution $\phi_{z_{dn}}$.

MGCTM. MGCTM extends LDA by dividing topics into two categories. One is the global topics used to capture corpus-level common semantics; the other is the local topics used to capture document-level specific semantics. Under MGCTM, there exists R global topics, and local topics are organized into J K-sized latent groups. To generate a document d, it (1) draw a multinomial distribution θ_d^g over global topics, from the global Dirichlet prior α^g; (2) choose a group η_d and draw a multinomial distribution θ_d^l over local topics, from the selected group-specific Dirichlet prior $\alpha_{\eta_d}^l$; (3) draw a Bernoulli decision distribution ω_d from the Dirichlet prior γ. To generate a word token, first sample a binary variable δ_{dn} from the Bernoulli decision distribution ω_d; If $\delta_{dn} = 1$, this word token w_{dn} will be generated from the global topic mixture proportion θ_d^g, otherwise w_{dn} will be generated from the local topic mixture proportion θ_d^l. The subsequent word generative process is the same as LDA. Profiting from the global topic design, MGCTM can uncover the common semantics and filter out the noise words (e.g., stopword-like words) in some degree.

Topic coherence. Topic coherence is a very popular automatical metric to evaluate whether the topics learnt by topic models are coherent. The intuition behind this metric is that a topic is more coherent if its most representative words are more frequently co-occurring. Given the top-M word list $V^k = \left(v_1^k, \ldots, v_M^k\right)$ in the topic distribution ϕ_k, the coherence value of this topic k is computed by:

$$Coh(k, V^k) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D\left(v_m^k, v_l^k\right) + \varepsilon}{D\left(v_l^k\right)} \tag{1}$$

where $D(v)$ is the number of documents containing the word type v; $D(v_1, v_2)$ is the number of documents containing both word type v_1 and v_2; ε is a smoothing constant used to avoid log zero. For topic coherence, values closer to zero imply greater co-occurrence, i.e., better coherent performance.

2.2. Problem

The story begins with an experiment on Twitter, where our original intention is to investigate topic modeling on short texts. Short texts contain very a few word tokens and are commonly quite noisy [16]. Documents from Twitter are typical short texts. We collect 1,000,000 Twitter documents from the web[1]; and then fit LDA and MGCTM[2] on this Twitter collection respectively in order to evaluate whether MGCTM can filter out the noise words in short texts. Surprisingly, we observe a very interesting result. Table 1 presents two top-10 highest probability words of corresponding topics learnt by LDA and MGCTM respectively. Their topic coherence values are listed in the first column. We observe that in contrast to LDA, MGCTM can effectively filter out some noise words (e.g., “lol” and “rt” in box[3]), but its topic coherence values are even worse than LDA’s (e.g., −256.5 vs. −255.7). This leads to an obvious conflict between human knowledge and the topic coherence metric.

Reviewing the topic word lists learnt by LDA and MGCTM in Table 1, we find that the only difference between corresponding

[1] After standard per-processing such as removal of stopwords, we obtain a Twitter collection with a 95,623-sized vocabulary.

[2] For LDA, the number of topics is set to 50; for MGCTM, the number of latent groups is set to 25, where each group contains 2 local topics, and the number of global topics is set to 5.

[3] The word “lol” is the abbreviation of “laugh out loud”, used as a modal particle; and the word “rt” is the abbreviation of “retweet”. Because they are too common in Twitter and lack specific semantics, we consider them as noise.

Table 2
Two top-10 word lists of corresponding topics learnt by LDA on S-Ng and NS-Ng, respectively. The first and third rows are topics learnt on S-Ng; the second and fourth rows are topics learnt on NS-Ng. The first column shows topic coherence values.

Topic coherence	Top-10 word list
−66.3	the, of, space, you, are, to, on, for, nasa, and
−179.6	space, nasa, earth, launch, gov, orbit, moon, shuttle, satellite, mission
−89.8	of, to, and, a, is, medical, for, disease, with, in
−188.6	health, medical, insurance, disease, doctor, treatment, patients, care, medicine, drug

topics is the two noise words “lol” and “rt”. Statistics results[4] show that both of them are much more frequently occurring than other top words, so they are more co-occurring with other top words in general, resulting in better topic coherence values. Based on this analysis, we believe that the conflict in topic coherence is caused by this kind of frequently occurring but meaningless words (i.e., just like the stopwords), existing in the top word list.

To further support our analysis, we conduct an additional experiment on the *Newsgroup* collection (See more detail about this collection in Section 4). We prepare two versions of *Newsgroup*, where one is the original collection (abbr. S-Ng) and the other is a processed collection (abbr. NS-Ng) by removing the stopwords. We simultaneously fit 50-topic LDA models on S-Ng and NS-Ng, and present two top-10 highest probability words of corresponding topics in Table 2. Unsurprisingly, we still see the conflict in the topic coherence metric mentioned above. Because the stopwords are much more frequently occurring, they dominate the top word lists learnt across S-Ng. Although they are much more co-occurring from each other, resulting in better topic coherence values, such top word lists filled with stopwords are useless.

In summary, the problems that we find are outlined as follows:

1. The standard top-M words ranked by the topic-word distributions ϕ are not always the most representative words.
2. The first problem above leads to a conflict problem between human knowledge and the topic coherence metric.

3. Methodology

Before introducing methodologies, we first propose basic settings and symbol definitions. In this work, we investigate novel topic representations around LDA and use collapsed Gibbs sampling (CGS) [17] for approximate LDA posterior inference. CGS involves sequentially resampling each topic assignment z_{dn} from its conditional posterior, holding all other variables fixed. Given final samples of z_{dn}, the point estimates of the topic-word distributions ϕ can be computed by:

$$\phi_{kv} = \frac{N_{kv} + \beta}{N_k + V\beta} \tag{2}$$

where N_{kv} and N_k are the number of the word type v assigned to the topic k and the total number of word tokens assigned to the topic k, respectively; V is the number of word types. Moreover, Let N_v be the number of the word type v has occurred; N be the total number of word tokens have occurred in a collection.

In terms of Eq. (2), given a topic k, ranking top-M words by the topic distribution ϕ_k is equivalent to finding M words with the largest N_{kv} values. This mechanism favors high frequency words, i.e., words with large N_v values. Frequently occurring but meaningless words, e.g., stopwords, sometimes mingle in the top-M word list, resulting in bad topic representation.

We consider that for a given topic k, a representative word v should be not only (1) with high marginal probability (i.e, large ϕ_{kv} values), but also (2) with low marginal probabilities in every other topic (i.e., small ϕ_{tv} values, where $t \in \{1, 2, \dots, K\}_{\neg k}$). The standard top-$M$ word list representation, ranked by the topic distributions ϕ, neglects the second factor. To improve it, we further consider the second factor to rerank topic words for more representative words. We propose three reranking methodologies in this paper, using (1) standard deviation weight (SDW), (2) standard deviation weight with topic size (SDWTS) and (3) Chi Square χ^2 statistic selection (CHI).

SDW reranking. Given a topic k, we provide weighting values for its distribution ϕ_k over all V words by:

$$weight_{SDW}(k, v) = \sqrt{\sum_{i \neq k} (\phi_{kv} - \phi_{iv})^2} \tag{3}$$

This weight is a pseudo standard deviation. It treats the processed partner (k, v), i.e., ϕ_{kv}, as the expectation, and computes its standard deviation over the probabilities of the same word v in every other topic. The intuition behind is that if a word v with larger ϕ_{kv} value in the topic k also corresponds to larger ϕ_{iv} values in most of other topics (i.e., $i \neq k$), we will provide a small weight to ϕ_{kv}; otherwise, we will provide a large weight to ϕ_{kv}. Finally, we can use the following weighted ϕ_{kv} values to rerank words in topics:

$$\phi_{kv}^{SDW} = weight_{SDW}(k, v) \times \phi_{kv} \tag{4}$$

SDWTS reranking. Based on SDW reranking, we further consider the topic size N_k (i.e., the number of word tokens assigned to each topic by Gibbs sampling). The topic size itself is a reasonable predictor of topic quality [11], where larger topic size implies better topic quality. Considering this, we present a novel weighting equation as follows:

$$\begin{aligned} weight_{SDWTS}(k, v) &= \sqrt{\sum_{i \neq k} (\phi_{kv} N_k - \phi_{iv} N_i)^2} \\ &\approx \sqrt{\sum_{i \neq k} (N_{kv} - N_{iv})^2} \end{aligned} \tag{5}$$

Finally, we can use the following weighted ϕ_{kv} values to rerank words in topics:

$$\phi_{kv}^{SDWTS} = weight_{SDWTS}(k, v) \times \phi_{kv} \tag{6}$$

Reviewing Eq. (5), we note that the SDWTS weight is focusing on the number of word tokens assigned to each topic directly.

CHI reranking. Chi Square χ^2 statistic has been widely used in feature selection for classification. It selects most discriminative features by measuring the statistical dependency between the feature and the category. In our case, we can treat words/topics as features/categories, and then directly use Chi Square χ^2 statistic to rank words in topics. The χ^2 with V different words and K topics is defined as:

$$\chi^2 = \sum_{k=1}^{K} \sum_{v=1}^{V} \frac{(N_{kv} - E_{kv})^2}{E_{kv}} \tag{7}$$

[4] The average document length of *Twitter* is only 4.2, however, “lol” occurs in 100K/1000K texts and “rt” occurs in 80K/1000K texts.

Table 3
Top-20 word lists learnt by LDA on NS-Ng. The first row is the topic about "food" and the second row is the topic about "baseball game".

NORM	SDW	SDWTS	CHI
food, gov, nasa, writes, eat, article, foods, don, apr, fat, chinese, sensitivity, brain, taste, jpl, eating, people, reaction, related, superstition	food, foods, eat, sensitivity, chinese, fat, taste, gov, restaurant, vegetables, nasa, eating, glutamate, dealy, meat, larc, milk, gsfc, allergic, brain	food, foods, eat, sensitivity, chinese, fat, taste, gov, restaurant, vegetables, nasa, eating, glutamate, dealy, meat, larc, milk, gsfc, allergic, elroy	food, foods, eat, chinese, eating, sensitivity, fat, taste, gov, restaurant, glutamate, dealy, meat, milk, allergic, brain, seizure, sugar, vegetables. diet
year, game, baseball, games, players, team, articles, runs, season, ride, technology, league, player, average, don, play, apr, pitcher, pitching, sox	baseball, game, players, braves, pitcher, pitching, sox, year, cubs, runs, games, jays, mets, season, team, phillies, pitchers, morris, alomar, pitch	baseball, game, players, braves, pitcher, pitching, innings, year, cubs, runs, games, jays, mets, season, team, phillies, pitchers, sox, alomar, pitch	baseball, game, players, braves, pitcher, pitching, sox, cubs, runs, year, games, jays, mets, season, phillies, pitchers, morris, alomar, pitch, team

Table 4
Top-20 word lists learnt by LDA on S-Ng. The first row is the topic about "food" and the second row is the topic about "baseball game".

NORM	SDW	SDWTS	CHI
the, and, to, i, that, of, it, a, is, in, t, msg, you, have, food, these, or, this, not, are	food, chinese, sensitivity, superstition, foods, taste, spdcc, restaurant, dyer, glutamate, eat, meat, cousineau, nmm, questor, compdyn, allergic, dougb, olney, sugar,	food, chinese, sensitivity, superstition, foods, taste, carcinogenic, restaurant, dyer, glutamate, eat, meat, cousineau, nmm, questor, compdyn, allergic, dougb, olney, blah,	food, chinese, sensitivity, superstition, foods, taste, spdcc, restaurant, dyer, glutamate, eat, meat, cousineau, nmm, allergic, questor, compdyn, dougb, olney, carcinogenic
in, a, i, s, to, edu, is, that, they, year, game, baseball, but, hit, games, writes, this, will, be, and	baseball, braves, hit, sox, magnus, gant, games, ohio, game, yankees, acs, reds, hitting, clutch, players, year, mets, dodgers, catcher, phillies	baseball, braves, hit, sox, magnus, gant, games, ohio, game, yankees, acs, reds, hitting, clutch, players, year, mets, dodgers, catcher, phillies	baseball, braves, hit, sox, magnus, gant, games, ohio, game, yankees, reds, acs, hitting, clutch, players, year, mets, dodgers, catcher, fans

where $E_{kv} = N_k N_v / N$. The χ^2 value for the word v in the topic k can be interpreted by the following probability:

$$\chi^2(k, v) = \frac{N\big(p(k, v)p(\neg k, \neg v) - p(\neg k, v)p(k, \neg v)\big)^2}{p(k)p(\neg k)p(v)p(\neg v)} \tag{8}$$

where $p(k, v)$ is the probability of the topic k containing the word v and $p(\neg k, \neg v)$ is the probability of not being in the topic k and not containing the word v and so on. Given all χ^2 values, we can rerank the vocabulary for every topic.

4. Evaluation

In this section, we evaluate our novel topic representations on two real-world collections. The first is *Newsgroup*[5], a collection of news releases, which contains 18,821 documents with a 93,864-sized vocabulary. We generate two versions of *Newsgroup*, where one is the original collection (abbr. S-Ng) and the other is a processed collection (abbr. NS-Ng) by removing the stopwords. The second collection is *Wikipedia*[6] (abbr. Wiki), which contains 1918 documents. After pre-processing, i.e., word stemming and removal of the stopwords, 9144 word types are left.

In our experiments, Gibbs sampler LDA is used to output the topic-word distributions ϕ. During model training, the iteration number of Gibbs sampler is set to 1000, and the Dirichlet hyperparameters α and β are set to 0.1 and 0.01, respectively. The top-M word list ranked by ϕ is called as NORM representation, and the top-M word list reranked by SDW/SDWTS/CHI methodology is called as SDW/SDWTS/CHI representation.

4.1. Evaluation on stopwords filtering

The first evaluation is on whether our reranking representations can filter out stopword-like words in the top list (i.e., top words occurring in most of the topics). For this goal, we fit a 100-topic LDA model on NS-Ng, and randomly select two topics, which are about food and baseball game, for visualization. The top-20 word lists of all four topic representations are illustrated in Table 3.

The first row is the top word list about food. Overall, we see that all four representations are in the main coherent. The word "food" is always the best representative word. However, in the NORM representation, some exalted words, such as "nasa" and "writes", are less related to the topic about food, and moreover, some words, such as "don" and "articles", seem the stopword-like words. The word "don" is obviously stopword-like. It is the first half of the stopword "don't", but is lucky to be left due to the misrecognition of the right single quote. We are a little surprised to judge that the word "articles" is stopword-like, because it has occurred in 35 top lists out of 100 topics, much more frequent than most of other top words. Compared to the NORM representation, the top word lists of the three reranking representations seem better. Top words in them are more prominent and precise about food. More importantly, they effectively filter out the stopword-like words, such as "don" and "articles". Besides, we see that the three reranking representations are almost the same, e.g., the only difference between SDW and SDWTS is the 20th word.

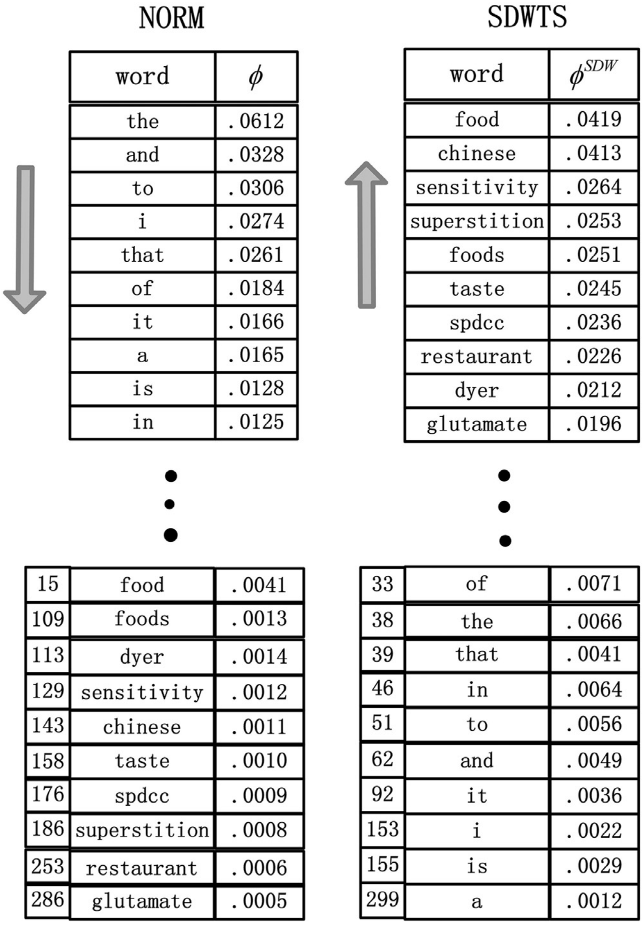
The second row is the top word list about the baseball game. In the NORM representation, again we see that the top words include some less relevant words, such as "ride" and "technology", and the two stopword-like words, i.e., "don" and "articles". Besides, its top one word "year" is less representative for this topic. In contrast, the three reranking representations seem better. First, they all successfully filter out the two stopword-like words. Second, they are obviously more coherent about the subject baseball game. Their top one and top two words maintain "baseball" and "game"; and other top words revolve around this subject.

We are further interested in whether our reranking representations can filter out the true stopwords in the top word list. For this purpose, we fit a 100-topic LDA model on S-Ng, and still visualize the two topics about food and baseball game. The top-20 word lists of all four topic representations are illustrated in Table 4.

We can observe that in the NORM representation, the top words are almost the stopwords, e.g., the topic about food contains 18/20 stopwords and the topic about baseball game contains 13/20 stopwords. Such kind of topics must be useless. In contrast, our reranking representations perform significantly better. They successfully filter out even the true stopwords, which are much frequently occurring in every document. Besides, we see that all three reranking representations seem coherent. Only a few less

[5] http://web.ist.utl.pt/acardoso/datasets/.

[6] http://topics.cs.princeton.edu/nubbi/nubbi_data_censored.tar.gz.



NORM

word	ϕ
the	.0612
and	.0328
to	.0306
i	.0274
that	.0261
of	.0184
it	.0166
a	.0165
is	.0128
in	.0125

15	food	.0041
109	foods	.0013
113	dyer	.0014
129	sensitivity	.0012
143	chinese	.0011
158	taste	.0010
176	spdcc	.0009
186	superstition	.0008
253	restaurant	.0006
286	glutamate	.0005

SDWTS

word	ϕ^{SDW}
food	.0419
chinese	.0413
sensitivity	.0264
superstition	.0253
foods	.0251
taste	.0245
spdcc	.0236
restaurant	.0226
dyer	.0212
glutamate	.0196

33	of	.0071
38	the	.0066
39	that	.0041
46	in	.0064
51	to	.0056
62	and	.0049
92	it	.0036
153	i	.0022
155	is	.0029
299	a	.0012

Fig. 1. The reranking processing of the SDWTS representation for the topic about food. The first/four column in the bottom is the word order in the NORM/SDWTS representation.

relevant words exist in the range from the 10th to 20th top words, e.g., “questor” and “nmm” in the topic about food, and “yankees” and “acs” in the topic about baseball game. Moreover, the top one words in reranking representations are representative for the corresponding topic, e.g., for the topic about food, we see that the top one word is just the word “food”.

Fig. 1 lists the top-10 words in the topic about food. Because the stopwords are with higher ϕ values, they dominate the top-10 list. The more representative words are lower ranked, e.g., the order of the word “foods” is only 109. The SDWTS reranking successfully filters out the stopwords from the top-10 list by providing larger weights to the dedicated words. We can observe that the ϕ^{SDWTS} values of the stopwords are much smaller than those of the final top words in the SDWTS representation.

4.2. Evaluation on human-interpretability

The second evaluation is on whether the top words in our reranking representations are more representative than those in the NORM representation. Because there are no gold-standard top word lists of topics, we use the word intrusion task [10] for indirect evaluation. In this task, each topic is presented by the top six words. Randomly remove one of six top words and then randomly select an intruder word to replace the removed top word. The task of users is asked to identify the intruder word. We consider that if the top topic words are more representative, it should be easier for users to identify the intruder word. Following this, for the same topic, more representative top words must lead to higher intrusion accuracy given by:

$$AC = \frac{\sum_{k=1}^{K} \mathbb{1}(i_k = w_k)}{K} \tag{9}$$

where i_k is the intruder word of the topic k selected by a human being, and w_k is the true intruder word of the topic k. In our experiments, we use an automatical intruder detector [13], which can emulate the performance of human judgements. Three different patterns are used to generate the intruder words: selecting (1) from the vocabulary (S_VOC), (2) from the top six words of other topics (S_TOPIC) and (3) from the 11th to 100th words of the current topic (S_SELF).

We fit a 150-topic LDA model on NS-Ng and a 100-topic LDA model on Wiki. See the following discussions on results of S_VOC, S_TOPIC and S_SELF tests.

Performance on S_VOC and S_TOPIC tests. For both tests, we randomly generate the intrusion topics 10 times, and finally present the average intrusion accuracy.

The performance on S_VOC test is shown in Fig. 2. For NS-Ng, the NORM representation accuracy is about 0.8, and the three reranking representations are all above 0.9. For Wiki, accuracy values of all the four are over 0.93, and reranking representations perform better, i.e., about 0.97. Among the three, the SDWTS and CHI representations perform better than the SDW representation and are more stable.

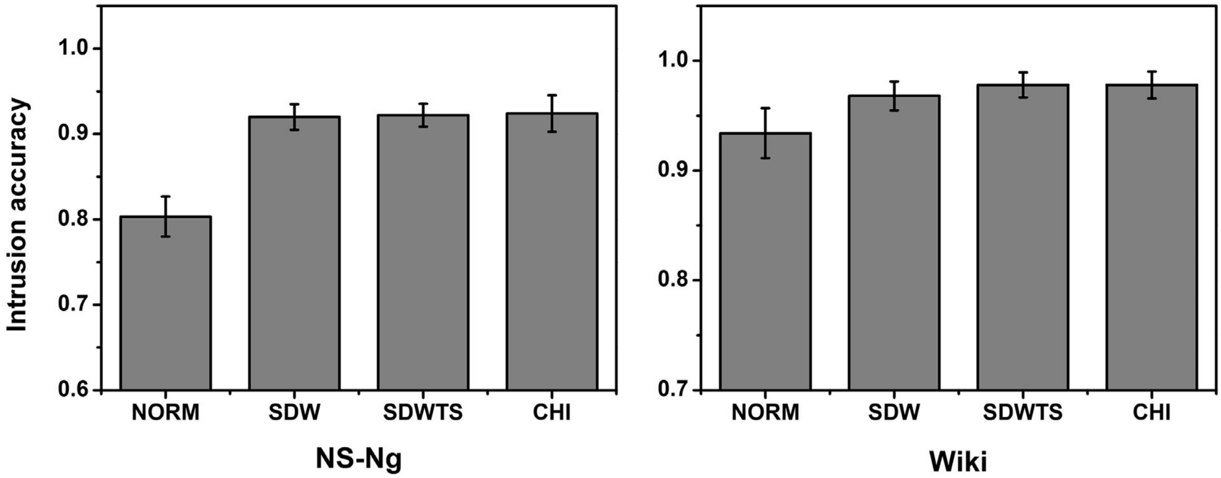


Fig. 2. Performance of S_VOC test on NS-Ng (left) and Wiki (right).

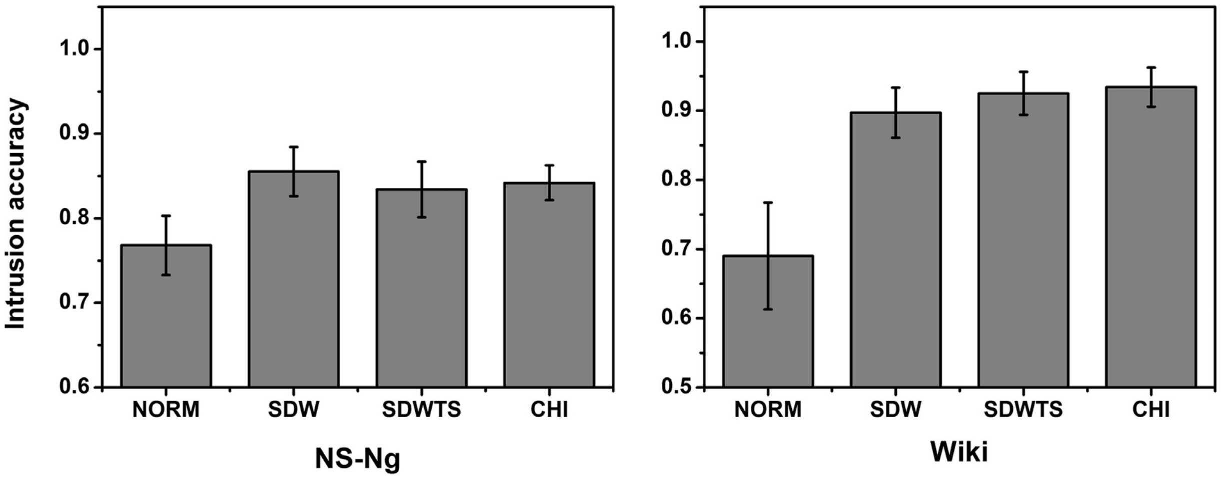


Fig. 3. Performance of S_TOPIC test on NS-Ng (left) and Wiki (right).

Table 5
The S_TOPIC intrusion task details of two lists of Wiki topics for all four topic representations. The words in bold are the true intruder words; and the words in box are the intruder words selected by the automatical intruder detector [13].

Representation	Top word list	Top word list
NORM	$\boxed{\text{don}}$ **rusian** home told father house	$\boxed{\text{school}}$ world america students americans **annotate**
SDW	started $\boxed{\textbf{rusian}}$ told wife father bed	school students schools $\boxed{\textbf{annotate}}$ america education
SDWTS	started home told $\boxed{\textbf{rusian}}$ father bed	school students schools movement $\boxed{\textbf{annotate}}$ education
CHI	started home told $\boxed{\textbf{rusian}}$ father bed	$\boxed{\textbf{annotate}}$ schools movement america education

The performance on S_TOPIC test is shown in Fig. 3. Overall, we can see that the three reranking representations perform significantly better than the NORM representation. For NS-Ng, the accuracy values of SDW and CHI representations are about 0.85, the SDWTS accuracy is about 0.83, but the NORM accuracy is just about 0.77. For Wiki, the performance gap is even larger, e.g., the accuracy values of the SDW and SDWTS representations are about 0.9, but the NORM accuracy is even lower than 0.7. Among the three reranking representations, the CHI accuracy values are the best. We see that the performance of S_TOPIC test is worse than that of S_VOC, especially for the NORM representation. We consider that this result is caused by two reasons. First, the NORM representation contains some high frequency but less representative words. These words seem intruder words themselves. Second, the NORM representation contains many stopword-like words, which are occurring in most of the top word lists. If such a word is selected as an intruder word to another topic that also contains this word. It must be difficult to identify the true intruder word. Table 5 shows two examples that the NORM representation loses. In the first example, the top word "don" looks more ecdemic than the true intruder word "rusian". In the second example, the intruder word "annotate" randomly selected from other topics is just the replaced top word in that topic. That is, the top topic words remain unchanged, resulting in difficulty in intruder identification. In contrast, the top words in our reranking representations are obviously more coherent. They have no meaningless words just like "don". We are thus concluding that our reranking representations are more representative.

Performance on S_SELF test. In this test we select intruder words from sub-top (from 11th to 100th) words in the current topic. Fig. 4 shows the average performance (on NS-Ng) of every ten adjacent different sub-top words, which are used as intruder words. Overall, we can see that our reranking representations perform a little better than the NORM representation, e.g., 0.61 (NORM) vs. 0.65 (SDW) on the 91th to 100th sub-top intruder words. This is another evidence that our reranking methodologies can top more representative words.

5. Related work

In topic modeling research, the topic is the word mixture proportion in form. How to express topics is an active direction. Basically, topics are represented as lists of top-M words with highest marginal probabilities. In some early researches, topics are manually labeled for convenient presentation of research results [18]. Recent attempts on topic representation are focusing on automatical topic labeling algorithms [5–7,19–21]. Most of them are based on the standard top-M word list representation, and generate candidate topic labels using external knowledge resources. Some typical works are illustrated as follows: automatical topic labeling proposed in [6] generates candidate labels from Wikipedia. It first uses the top-10 topic words to query Wikipedia for relevant article titles, and then uses these titles to generate secondary candidate labels. Finally, the candidate labels are ranked by a supervised model. The authors of [7] propose an algorithm to label topics by images, instead of text. This algorithm collects candidate image labels by querying Google with top-5 topic words, where the search is restricted to the English Wikipedia. Textual information from the metadata and visual features (e.g., SIFT descriptor) extracted from images [22–26] are used in ranking the candidate image labels. Besides, we see an interesting comparison among different topic representations within a document retrieval task [21].

Nowadays, mainstream topic representations are based on the standard top-M word list, and moreover, the automatical topic quality metrics [10–12] are also based on it. Compared with the standard top-M word list, our reranking methodologies further take this factor into consideration. The reranking top-M words are,

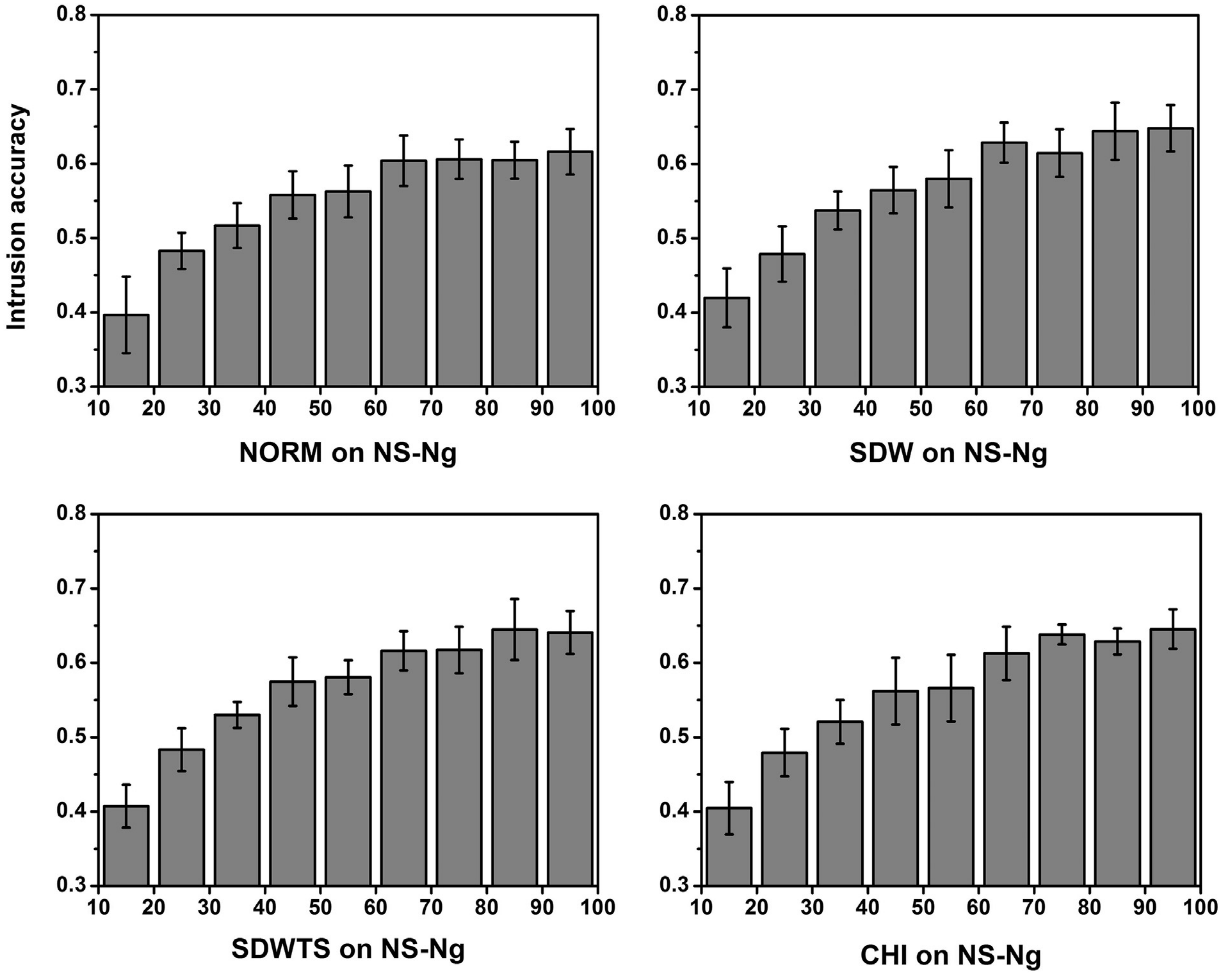


Fig. 4. Performance of S_SELF test on NS-Ng.

empirically, much more representative for a given topic. The automatical topic labeling in [19] also follows a reranking idea. However, it only reranks the topic words in the top-*M* list. In contrast, our methodologies rerank the vocabulary.

6. Conclusion

We investigate how to find more representative word lists to represent topics learnt by topic models. We propose three methodologies to rerank topic words by considering the marginal probabilities on words over different topics. The reranked top-*M* words are used as novel topic representations, namely SDW, SDWTS and CHI representations. Experimental results indicate that our methodologies can (1) effectively filter out stopword-like words and (2) find more representative topic words comparing with the standard topic words with highest marginal probability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) [grant numbers 61876071, 61602204], and the State Key Laboratory of applied optics.

References

[1] D.M. Blei, Probabilistic topic models, Commun. ACM 55 (4) (2012) 77–84.
[2] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[3] A. Haghighi, L. Vanderwende, Exploring content models for multi-document summarization, in: North American Chapter of the Association for Computational LinguisticsCHuman Language Technologies, 2009, pp. 362–370.
[4] J.H. Lau, P. Cook, D. McCarthy, D. Newman, T. Baldwin, Word sense induction for novel sense detection, in: Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 591–601.
[5] Q. Mei, X. Shen, C. Zhain, Automatic labeling of multinomial topic models, in: International Conference on Knowledge Discovery and Data Minings, 2007, pp. 490–499.
[6] J.H. Lau, K. Grieser, D. Newman, T. Baldwin, Automatic labelling of topic models, in: Annual Meeting of the Association for Computational Linguistics, 2011, pp. 1536–1545.
[7] N. Aletras, M. Stevenson, Representing topics using images, in: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 158–167.
[8] Q. Castellà, C.A. Sutton, Word storms: multiples of word clouds for visual comparison of documents, in: 23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, 2014, pp. 665–676.
[9] L.v.d. Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.
[10] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, D.M. Blei, Reading tea leaves: How humans interpret topic models, Neural Information Processing Systems, 2009.
[11] D. Mimno, H.M. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: Conference on Empirical Methods in Natural Language Processing, 2011.
[12] N. Aletras, M. Stevensonm, Evaluating topic coherence using distributional semanticss, in: International Conference on Computational Semantics, 2013, pp. 13–22.
[13] J.H. Lau, D. Newman, T. Baldwin, Machine reading tea leaves: automatically evaluating topic coherence and topic model quality, in: Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 530–539.
[14] P. Xie, E. P.Xing, Integrating document clustering and topic modeling, in: International Conference on Conference on Uncertainty in Artificial Intelligence, 2013.

[15] X. Li, A. Zhang, C. Li, J. Ouyang, Y. Cai, Exploring coherent topics by topic modeling with term weighting, Inf. Process. Manage. (2018).

[16] X. Li, Y. Wang, A. Zhang, C. Li, J. Chi, J. Ouyang, Filtering out the noise in short text topic modeling, Inf. Sci. 456 (2018) 83–96.

[17] T.L. Griffiths, M. Steyvers, Finding scientific topics, Natl. Acad. Sci. 101 (suppl. 1) (2004) 5228–5235.

[18] Q. Mei, C. Zhai, Discovering evolutionary theme patterns from text: an exploration of temporal text mining, in: ACM International Conference on Knowledge Discovery and Data Mining, 2005, pp. 198–207.

[19] J.H. Lau, D. Newman, S. Karimi, T. Baldwin, Best topicword selection for topic labelling, in: International Conference on Computational Linguistics, 2010, pp. 605–613.

[20] N. Aletras, M. Stevenson, Labelling topics using unsupervised graph-based methods, in: Annual Meeting of the Association for Computational Linguistics, 2014, pp. 631–636.

[21] N. Aletras, T. Baldwin, J.H. Lau, M. Stevensoni, Evaluating topic representations for exploring document collectionss, J. Assoc. Inf. Sci. Technol. (2015) 1.

[22] L. Wu, Y. Wang, J. Gao, X. Li, Deep adaptive feature embedding with local sample distributions for person re-identification, Pattern Recognit. 73 (2018) 275–288.

[23] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, X. Huang, Robust subspace clustering for multi-view data by exploiting correlation consensus, IEEE Trans. Image Process. 24 (11) (2015) 3939–3949.

[24] Y. Wang, W. Zhang, L. Wu, X. Lin, M. Fang, S. Pan, Iterative views agreement: an iterative low-rank based structured optimization method to multi-view spectral clustering, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, 2016, pp. 2153–2159.

[25] Y. Wang, X. Lin, L. Wu, W. Zhang, Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval, IEEE Trans. Image Process. 26 (3) (2017) 1393–1404.

[26] Y. Wang, W. Zhang, L. Wu, X. Lin, X. Zhao, Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion, IEEE Trans. Neural Netw. Learn. Syst. 28 (1) (2017) 57–70.