Information Sciences 519 (2020) 243–259

Contents lists available at ScienceDirect

Information Sciences

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

An association-constrained LDA model for joint extraction of product aspects and opinions

Changxuan Wan[a,*], Yun Peng[b], Keli Xiao[c,*], Xiping Liu[a], Tengjiao Jiang[a], Dexi Liu[a]

[a] School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China
[b] School of Computer Information Engineering, Jiangxi Normal University, Nanchang, China
[c] College of Business, Stony Brook University, USA

ARTICLE INFO

Article history:
Received 13 April 2019
Revised 25 November 2019
Accepted 17 January 2020
Available online 18 January 2020

Keywords:
Aspect words
Association constraint
LDA model
Opinion words

ABSTRACT

The Latent Dirichlet Allocation (LDA) model, which is a document-level probabilistic model, has been widely used in topic modeling. However, an essential issue of the LDA is its shortage in identifying co-occurrence relationships (e.g., aspect-aspect, aspect-opinion, etc.) in sentences. To address the problem, we propose an association constrained LDA (AC-LDA) for effectively capturing the co-occurrence relationships. Specifically, based on the basic features of the syntactic structure in product reviews, we formalize three major types of word association combinations and then carefully design corresponding identifications. For reducing the influence of global aspect words on the local distribution, we apply an important constraint on global aspects. Finally, the constraint and related association combinations are merged into the LDA to guide the topic-words allocation in the learning process. Based on the experiments on real-world product review data, we demonstrate that our model can effectively capture the relationships hidden in local sentences and further increase the extraction rate of fine-grained aspects and opinion words. Our results confirm the superiority of the AC-LDA over the state-of-the-art methods in terms of the extraction accuracy. We also verify the strength of our method in identifying irregularly appeared terms, such as non-aspect opinions, low-frequency words, and secondary aspects.

1. Introduction

Sentiment analysis has been widely applied to product reviews for a variety of applications, such as supporting marketing strategies, customer analysis, etc. The initial and fundamental step of sentiment analysis is to identify and extract keywords expressing product aspects and related opinions. We consider two types of words/terms, global and local ones, which represent global and local aspects with corresponding opinions, respectively. A global aspect is the general subject of a product or service, such as “camera”, “seller attitude”, “quality”, etc.; a local aspect is an attribute or a component of a global aspect, such as “screen”, “battery”, “price”, etc. Regarding expressed opinions on reviews, global opinions describe opinions toward global aspects, such as “good”, “not bad”, etc.; and local opinions refer to those describing local aspects, such as “cheap”, “clear”, etc. Generally, potential buyers view local aspects and related opinions as more important information regarding product details. Therefore, identifying and analyzing local aspects and opinions as well as their relationships

* Corresponding authors.
E-mail addresses: wancx@jxufe.edu.cn (C. Wan), keli.xiao@stonybrook.edu (K. Xiao).

https://doi.org/10.1016/j.ins.2020.01.036
0020-0255/

are non-trivial tasks in the *fine-grained sentiment analysis* [1]. For the tasks of aspect and opinion mining, the classic Latent Dirichlet Allocation (LDA) model assumes that a document is a random mixing of topics, and each topic can be considered as a specific probability distribution of words. This assumption can lead to a considerable reduction of the dimensionality of data and hence benefit the extraction for potential topical words covering the aspects and opinions.

However, considering that the standard LDA model is a document-level probability model, there are four important issues that have not been well addressed in existing methods when using it in fine-grained sentiment analysis. First, it is difficult for the LDA to extract opinion words when aspects are missing in a sentence. A common way of using LDA models to address the problem is to assign opinion words without aspects to a default topic consists of high-frequent aspect words. However, the process may mislead the opinion mining process. Second, it is difficult to use standard LDA models to extract low-frequency opinion words. Third, LDA can easily fail in secondary aspect identification. Generally, an aspect is usually an attribute of a product, while a second aspect is a subdivision of an attribute. A secondary aspect is usually a low-frequency word, which often appears with local aspect words. For instance, “durability” in “battery durability”, “sensitivity” in “touch-screen sensitivity”, and “dust" in “dust on the lens”. Last, since global aspect words may have high word frequency and high document frequency, they can be easily allocated to different topics. As a result, local aspect words with relatively low frequency would have a small chance to be extracted, while global aspects may be retrieved multiple times. Although models have been developed to extract aspect words and opinion words simultaneously (see [2,3] as examples), most of them failed to effectively consider the semantic association between words (e.g., aspect-aspect, aspect-opinion association, etc.). The shortage of semantic association information may lead to further failure in extracting low-frequency opinions and secondary aspects.

To this end, our work aims to fundamentally address the aforementioned issues and extract accurate aspects and opinions from texts of product reviews by extending and enhancing the LDA model. In particular, our proposed model improves the semantic understanding ability of topic models by using semantic associations among words as prior knowledge. By assigning the global aspect words to a small number of global topics through the global aspect constraint, our model can reduce the chance that global aspect words appear in local topics, and meanwhile, it enlarges the dissimilarity between global topics and local topics. Also, the new model effectively considers the semantic association information between words when assigning words to local topics, e.g., {(*local aspect word, low-frequency opinion word*)} and {(*local aspect word, secondary aspect word*)}. The semantic association information can increase the allocation weight of low-frequency opinion words and secondary aspect words under local topics, leading to a better performance on extracting low-frequency opinion words and secondary aspect words. Moreover, by mining missing aspect words that are most likely to be matched with the opinion words, our model offers better guidance to help decide the topic allocation of non-aspect opinion words. That is, if we name the identified missing aspect words as the *default aspect words*, our model is capable to mine the semantic association {(*default aspect word, non-aspect opinion word*)}, and hence the opinion words with missing aspect words can be effectively extracted.

To sum up, our paper contributes to the literature in four ways.

- We define three essential types of word association combinations, including: (i) non-aspect opinion words and their corresponding default aspect words, (ii) low-frequency opinion words and local aspect words, and (iii) secondary aspect words and local aspect words. We also propose discovery rules and the method of calculating the associative strengths. The rules consider the association combinations not only between aspects and opinions but also between aspects and aspects. Through the association constraints, the cohesion of the topic words can be largely improved.
- By formalizing a constraint, global aspect words can be distributed to a small number of topics, and their distribution probabilities under other topics can be reduced. The constraint can also reduce the impact of global aspect words on the local word distributions, and hence enhance the ability of our model for the identification of global topics and local topics.
- Different from classic LDA models which are only based on a priori probability distribution, we develop an Association Constrained LDA (AC-LDA) model. Our model allows the constraints on global aspect words and association combinations to influence the topic allocation.
- We conduct extensive experiments to verify the effectiveness of the proposed methods. The results show that the AC-LDA model outperforms existing forms of LDA in the global and local aspect extraction, as well as the coherence of aspects and opinions. The new method also increases the extraction rate of low-frequency words, aspects, and opinions, as well as non-aspect opinions.

The rest of the paper is organized as follows. Section 2 discusses the related work. Following Section 3, where association analysis is described, we introduce the AC-LDA model in Section 4. Then, in Section 5, we discuss the experiments and present the results. Finally, we conclude the paper in Section 6.

2. Related work

Related work on LDA-based methods falls into two categories: sole aspect extraction and joint extraction for aspects and opinions.

2.1. Sole aspect extraction

Following the standard LDA [4] model that aims to produce global topic words, there is an extensive literature on extending the LDA for extracting fine-grained topic words. For example, Titov et al. [5] proposed a multi-grain LDA by assuming that global topics capture the overall attributes of products, and local topics handle the fine-grained aspects appraised by users. The model can not only extract the aspects but also cluster similar aspects into the same topic. Following a similar idea, we propose the concept of global aspect constraint for allocating global aspect words into a small number of global topics. This design enlarges the distinction between global and local topics, and hence reduces the probability of assigning global aspect words into local topics.

By adding domain knowledge to LDA through a Dirichlet forest prior, Andrzejewski et al. [6] developed a Dirichlet forest LDA to capture language characteristics for different domains in topic modeling tasks. Specifically, the Dirichlet forest LDA includes *Cannot-Link* and *Must-Link* constraints as prior knowledge. The *Must-Link*(u, v) increases the probability for u and v to be allocated to the same topic, and the *Cannot-Link*(u, v) constraint suggests a low probability for the words to be allocated into the same topic. Related work with similar ideas can also be found in [7,8]. However, an important issue associated with the growth in the number of documents is the exponentially increased computational complexity. Thus, researchers seek other ways to include prior knowledge.

Aiming at addressing the inconsistency in aspect extraction based on unsupervised topic models, Chen et al. [9] proposed an automated knowledge LDA (AKL) model to guide the aspect extraction. The acquisition of prior knowledge is automatically obtained from the product reviews derived from different domains. The AKL model can deal with knowledge acquired from various fields and the model's fault tolerance of knowledge. Bagheri et al. [10] proposed an ADM-LDA (aspect detection model based on latent Dirichlet allocation) model to extract aspects from sentences. Different from the word bag of standard LDA, the ADM-LDA assumes that related aspect words in a sentence can be modeled based on a Markov chain.

A critical issue of existing LDA models as mentioned above for aspect extraction is that they do not consider the associations between aspects and opinions, which may significantly affect the topic allocation. Therefore, our work focuses on addressing the joint extraction task for aspects and opinions.

2.2. Joint extraction for aspects and opinions

To capture opinion information during the aspect extraction, Lin et al. [2] added the sentiment-level feature and considered sentiment distribution in their sentiment topic joint (JST) model. In the JST model, pre-defined opinion words were used as prior knowledge, and the knowledge was used to decide the sentiment label and the topic allocation of words in the initiation of topical post probability. Another example is The MaxEnt-LDA (Maximum Entropy LDA) model [11], which uses syntactic features to distinguish aspect words and opinion words. Also, Alam et al. [12] proposed a domain-independent topic sentiment model called Joint Multi-grain Topic Sentiment (JMTS) to extract sentiment-oriented ratable aspects. Lu et al. [13] proposed a Sentiment Topic Model (STM), which predicts the sentiment polarity through regression models. However, such prior knowledge ignores the correlation between aspects and opinions. Similar work can also be found in [14–16].

Assuming that a sentence has only one aspect, and all words in a sentence are generated by one aspect, Jo et al. [3] developed a Sentence-LDA (SLDA) for aspect extraction. ASUM (aspect and sentiment unification model), an extension of the SLDA model, was used to model the sentiments and aspects simultaneously. However, the model cannot consider word-level and sentence-level aspect-opinion associations. Mukherjee et al. [17] proposed two probabilistic models based on the seed words classification to extract and classify the aspect words automatically. The models could be used to model the aspect and opinion words in a better way. The first model was the SAS (Seeded Aspect and Sentiment) model, and the second ME-SAS (Maximum Entropy SAS) model was added to the maximum entropy prior to distinguish aspect words and opinion words. Chen et al. [18] proposed a topic model that automatically generated Must-links and Cannot-links (AMC). This model first mined some reliable (prior) knowledge from the past learning/modeling results and then used it to guide the model inference to generate more coherent topics. Heyrani-Nobari et al. [19] presented an unsupervised generative model AS-AC (aspect-action) LDA, to extract aspect-behavior pairs from online reviews. The model can capture aspects, behaviors as well as their relations based on an assumption that each user's comments were generated by mixing aspects and behaviors. That is, the model was designed to get the latent factors of aspect types and behavior intention, which described how users formed a topic in the reviews.

To identify product aspects more accurately, some models consider the co-occurrence relationship as prior domain knowledge, for instance, the ELDA (enriched LDA) model [20]. The ELDA can extract fundamental product aspects and meanwhile automatically extracts the prior knowledge from co-occurrence relationships and related topic aspects. Given that only the relationship of product subcategories is considered in the ELDA, and the parent categorization is neglected, a CAT-LDA (categorical LDA) model was developed by adding the subcategory relationship between the parent category and subcategory as prior knowledge [21]. Amplayo et al. [22] proposed a model called Micro Aspect Sentiment Model (MicroASM), focusing on aspect-level term extraction and document-level sentiment classification for short reviews. For improving the performance of aspect rating, Xue et al. [23] proposed a topic model that can explicitly estimate aspect ratings. The model merges the sentiment features of the modifier terms with those of the aspects. For the aspect and opinion extraction for product and service review, many models with similar ideas as discussed above were developed [24,25].

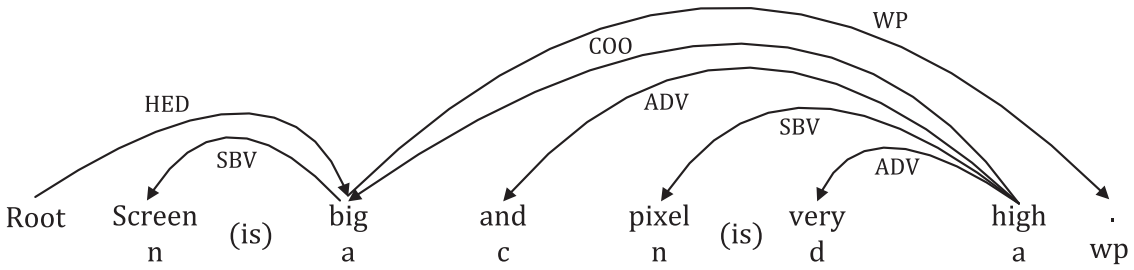


Fig. 1. The Dependency parsing and the POS tagging result of Example 1. *Note*: The POS tags "a", "n", "d", "c", and "wp" refer to adjective, noun, adverb, conjunction and punctuation, respectively. The dependency labels HED, SBV, ADV and COO indicate head, subject-verb, adverbial and coordinate, respectively.

To sum up, an important issue shared in existing LDA models is that they rarely consider the semantic relationship of words at the sentence-level. This situation is even more challenging when handling Chinese review texts. Therefore, our work focuses on addressing the joint extraction task for aspects and opinions by taking the sentence as the unit. We capture the associations between aspects and aspects, as well as the relationship between aspects and opinions. We enhance the LDA model by adding the obtained associations as prior and finally receive an improved cohesion and consistency of the words on the same topic.

3. Association combinations model and definitions

In this section, we introduce three types of word association combinations: non-aspect opinion words and their corresponding default aspects (*NA associations*), low-frequency opinion words and local aspect words (*LO associations*), and the secondary aspect words and the local aspect words (*SA associations*). This section also describes the rules to detect the association combinations and the method for computing-related associative strengths.

3.1. NA associations

In product reviews, many sentences contain non-aspect opinion words, such as "very fashionable" and "really good" etc. Although the corresponding aspect words are omitted in these sentences, they may appear in other places, such as "appearance is very fashionable" and "the camera is really good" etc. Therefore, the association combinations of aspect words and opinion words in other sentences may be used to identify aspects of the current non-aspect opinion words, and further affect the topic-words allocation. Note that evidence in support of the regular existence of NA associations can be found in [26], in which the linguistic concept of lexical collocation is introduced.

We use (*aspect word, opinion word*) to represent an *aspect-opinion* combination and propose a set of rules to locate it for handling non-aspect opinion words. Then, we introduce an equation calculating the associative strengths of *aspect-opinion* combinations. Given a non-aspect opinion word, the possible associated aspect word can be found from S_{ca1} the candidate set of *aspect-opinion* combinations. Finally, the set of *NA associations* is built as $S_{NA} = \{$(*default aspect word, non-aspect opinion word*)$\}$. We propose the following three-step process to obtain the *NA associations* and their associative strengths.

Step 1: We obtain the set of candidate aspect-opinion combinations.

To achieve the task, we formalize the following rule for aspect-opinion combination discovering, and then we provide an example to demonstrate the process.

Rule 1. If the POS of <*parent node, child node*> in dependency relation SBV (i.e., subject-verb) satisfies <adjective, noun>, then the noun acts as an aspect word, and the adjective acts as an opinion word.

Example 1. "The screen (is) big and the pixel (is) very high". The result of dependency parsing and POS tagging of the sentence is shown in Fig. 1. The sentence includes two *aspect-opinion* combinations: (screen, big) and (pixel, high).

In Fig. 1, letters at the bottom indicate the POS of words. Also, we represent a dependency relationship between words as: "*type of dependency* <*parent node*(POS), *child node*(POS)>". For example, SBV<big(a), screen(n)>. As can be seen, the aspects and opinions in *aspect-opinion* combinations in the sentence have obvious modification relationship, and they also have a fixed relationship in terms of POS, e.g., SBV<big(a), screen(n)> and SBV<high(a), pixel(n)>. Therefore, we can use the dependency parsing to find the typical sentence structure of *aspect-opinion* combinations. Then, we apply a dependency relation, denoted by *type of dependency* <adjective, noun>, to get a candidate *aspect-opinion* combination, denoted by (noun, adjective). We represent the set of these candidate *aspect-opinion* combinations as S_{ca1}.

Step 2: We identify non-aspect opinion words.

We formalize the following rule to obtain non-aspect opinion words and then provide an example to show the usage of rule.

Rule 2. If the POS of <*parent node*, *child node*> in dependency relation ADV (i.e., adverbial) satisfies <adjective, adverb> and the adjective is not the parent node of SBV, then the adjective acts as a non-aspect opinion word.

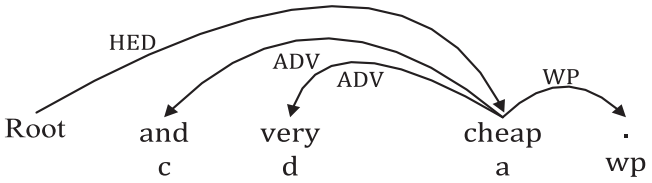


Fig. 2. The Dependency parsing and the POS tagging result of Example 2. *Note*: The POS tags "a", "d", "c", and "wp" refer to adjective, adverb, conjunction and punctuation, respectively. The dependency labels HED and ADV indicate head and adverbial, respectively.

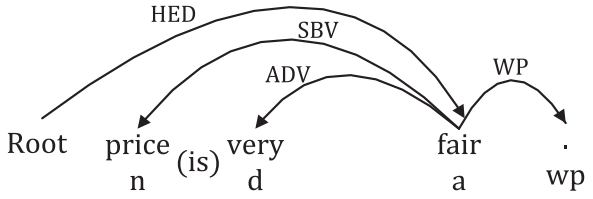


Fig. 3. The Dependency parsing and the POS tagging result of Example 3. *Note*: The POS tags "a", "n", "d", and "wp" refer to adjective, noun, adverb, and punctuation, respectively. The dependency labels HED, SBV, and ADV indicate head, subject-verb, and adverbial, respectively.

Example 2. "and very cheap". The result of dependency parsing and POS tagging of the sentence is shown in Fig. 2, and the "cheap" is a non-aspect opinion word.

From Fig. 2, it can be seen that there is no SBV in the sentence, but there exists dependency relation ADV<cheap(a), very(d)>. The "cheap" is a non-aspect opinion word in this sentence. It can be seen that the POS of "cheap" is an adjective and it is a parent node in ADV. Therefore, we can use the dependency parsing and POS to find non-aspect opinion words.

Step 3: We compute associative strengths and identify default aspect words.

We first define S-PMI (Sentence Pointwise Mutual Information) to measure the associative strength of an *aspect-opinion* combination (w_i, w_j) in S_{ca1}. Because the relationships between the aspects and the opinions are generally included in a single sentence, the co-occurrence frequency of words w_i and w_j in (w_i, w_j) is computed at the sentence level. S-PMI is defined as:

$$S{-}PMI(w_i, w_j) = \frac{\lg f_c(w_i, w_j)}{\lg f(w_i) \lg f(w_j)}, \tag{1}$$

where $f_c(w_i, w_j)$ is the sentence co-occurrence frequency of words w_i and w_j; and $f(w_i)$ and $f(w_j)$ are the frequencies of words w_i and w_j, respectively.

Then, given a non-aspect opinion word w_o and a candidate set S_{ca1} of *aspect-opinion* combinations, we find the aspect word w_a in S_{ca1} such that: (i) S-PMI (w_a, w_o) is largest; and (ii) w_a and w_o satisfy the Rule 2. Then w_a will be treated as the default aspect word associated with w_o. By this way, the set of *NA associations* can be constructed as S_{NA} ={(*default aspect word, non-aspect opinion word*)}.

3.2. LO associations

The syntactic structure and the POS of the low-frequency opinion words and the local aspect words also satisfy Rule 1, as shown in Fig. 3. Therefore, based on the set S_{ca1}, by redefining the equation calculating the associative strengths of the *aspect-opinion* combinations, we can obtain the set of *LO associations*, denoted by S_{LO}={(*local aspect word, low-frequency opinion word*)}.

Example 3. "The price (is) very fair". The result of dependency parsing and POS tagging of the sentence is shown in Fig. 3. The sentence includes one *aspect-opinion* combination (price, fair).

Because of the low co-occurrence of low-frequency opinion words and local aspect words, considering only the co-occurrence is not enough. For example, considering "appearance is very fashionable" and "price is very reasonable", in these sentences aspect words and opinion words have certain relationships, especially the low-frequency opinion words are generally used to modify specific aspect words. Some of the high-frequency opinion words also have similar features, such as "the price (is) very cheap", these high-frequency opinion words can be found by LDA directly. Therefore, when calculating the associative strengths of the *LO* associations, we consider not only the exclusion of co-occurrence but also the threshold limit of opinion word frequency.

To increase the associative strengths between low-frequency opinion words and their modified local aspect words, the frequency ratio and the co-occurrence frequency difference are introduced, and the co-occurrence frequency and mutual

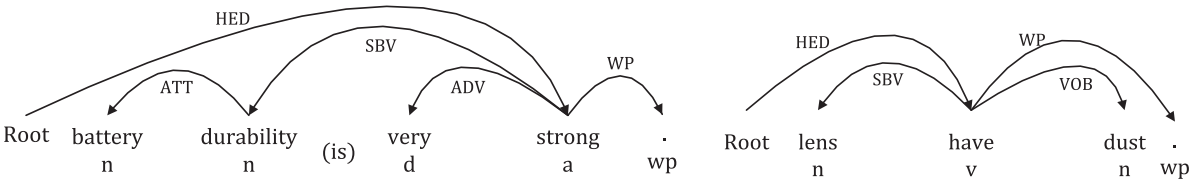


Fig. 4. The Dependency parsing and the POS tagging result of Example 4. *Note*: The POS tags "a", "n", "d", "v", and "wp" refer to adjective, noun, adverb, verb, and punctuation, respectively. The dependency labels HED, ATT, SBV, ADV, and VOB indicate head, attribute, subject-verb, adverbial, and verb-object, respectively.

exclusivity of the words are analyzed at the same time. The *LO* associative strength can be computed as:

$$R(w_i, w_j) = \begin{cases} \frac{\lg f_c(w_i, w_j) \lg p'}{\lg f'(w_i) \lg f'(w_j)} & f(w_j) \le \zeta_1 \\ 0 & f(w_j) > \zeta_1, \end{cases} \tag{2}$$

where ζ_1 is the word frequency threshold. $p' = \frac{f(w_i)}{f(w_j)}$, where $f(w_i)$ and $f(w_j)$ are frequencies of w_i and w_j respectively; the higher the value of p', the higher the associative degree of w_j on w_i; $f'(w_i)$ is the difference between word frequency $f(w_i)$ and co-occurrence frequency $f_c(w_i, w_j)$, the lower the value of $f'(w_i)$, the higher the exclusive degree of w_i on w_j.

Then, the normalized strength can be obtained as:

$$R(w_i, w_j) = \frac{R(w_i, w_j) - min_R}{max_R - min_R}, \tag{3}$$

where max_R is the maximum value of associative strengths, min_R is the minimum value of associative strengths.

After deleting the candidate *aspect-opinion* combinations whose associative strength is 0 from the set S_{ca1}, and we can obtain the set of *LO associations*, denoted as S_{LO} ={(*local aspect word, low-frequency opinion word*)}.

3.3. SA associations

To get the *SA associations*, we first set rules to obtain the candidate set of the combinations of local aspect words and secondary aspect words, S_{ca2}. The *aspect-aspect* combination is denoted by (*aspect word, aspect word*). Then, we define the equation for computing the associative strengths of the candidate *aspect-aspect* combinations. Finally, the set of *SA associations* is built, denoted as S_{SA} ={(*local aspect word, secondary aspect word*)}.

Step 1: Obtaining candidate *aspect-aspect* combinations.

We first formalize a rule to identify aspect-aspect combinations and then provide an example sentence to show the usage of the rule.

Rule 3. (a) If the POS of <*parent node, child node*> in dependency relation ATT (i.e., attribute) satisfies <noun, noun>, then the *parent node* in ATT acts as a secondary aspect and the *child node* in ATT acts as a local aspect. (b) If the POSs of *child nodes* in dependency relation SBV and VOB (i.e., verb-object) are all noun, and the *parent node* in SBV is equal to the *parent node* in VOB, then the *child node* in SBV acts as a local aspect and the *child node* in VOB acts as a secondary aspect.

Example 4. "Battery durability (is) very strong"; "Lens have dust". The results of dependency parsing and POS tagging of the two sentences are shown in Fig. 4. The two sentences include two *aspect-aspect* combinations: (battery, durability) and (lens, dust).

As we can see, the local aspects and secondary aspects that follow the *aspect-aspect* format in the sentences have obvious modification relationship, and they have fixed relationships of POSs. For example, "battery" is a local aspect, and "durability" is a secondary aspect. They satisfy ATT<*secondary aspect*(n), *local aspect*(n)>. "lens" is a local aspect, and "dust" is a secondary aspect. They satisfy SBV<*dependent verb*(v), *local aspect*(n)> and VOB<*dependent verb*(v), *secondary aspect*(n)>, respectively. Note that the dependent verb of SBV is equal to the dependent verb of VOB. Therefore, we can use the dependency parsing to find the sentence structure of these words and apply the POS relationship <noun, noun> to get the candidate set of *aspect-aspect* combinations, denoted as S_{ca2} = {(*local aspect word, secondary aspect word*)}.

Step 2: Calculating associative strengths and discovering association combinations.

Secondary aspect words are usually relatively low-frequency and often rely on local aspect words. For example, in "sensitivity of the screen is high", the secondary aspect "sensitivity" is dependent on the local aspect "screen". As a result, the associative strengths of these words should consider the co-occurrence exclusion. Similarly, non-secondary aspects also have similar dependencies, such as the "screen" is dependent on "camera" in "the camera screen is very clear", but the frequencies of non-secondary aspects are much higher compared with the secondary aspects. Thus, we introduce a frequency threshold when computing the associative strengths of these words. That is,

$$R(w_i, w_j) = \begin{cases} \frac{\lg f_c(w_i, w_j)}{\lg f(w_i) \lg f'(w_j)} & f(w_j) \le \zeta_2 \\ 0 & f(w_j) > \zeta_2, \end{cases} \tag{4}$$

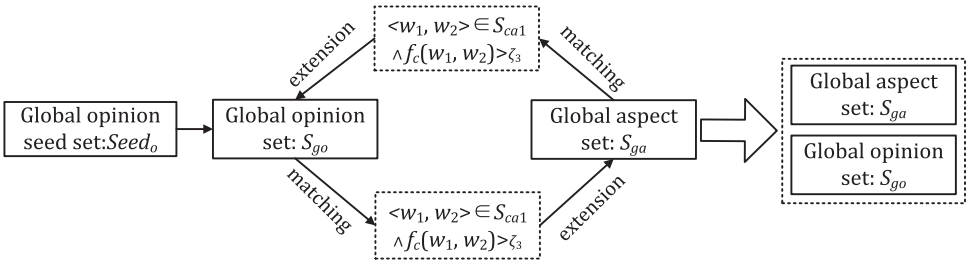


Fig. 5. Iterative discovery process of global aspects.

where ζ_2 is the word frequency threshold; $f'(w_j)$ is the difference between the word frequency $f(w_j)$ and the co-occurrence frequency $f_c(w_i, w_j)$, the lower the value of $f'(w_j)$. The higher the exclusive degree of w_j on w_i. As the secondary aspects are only associated with fixed local aspects, the frequencies of these words are almost equal to the co-occurrence frequencies of secondary aspects and local aspects.

After deleting the candidate *aspect-aspect* combinations whose associative strength is zero from the set S_{ca2}, we can obtain the set of *SA associations*, denoted by $S_{SA} = \{(\textit{local aspect word, secondary aspect word})\}$.

4. The AC-LDA model

This section introduces the AC-LDA (association constrained LDA) model, which incorporates the prior knowledge about the global aspect words and the captured association combinations.

Since the frequency of global aspect words is much higher than other aspects, it influences the topic allocation of the low-frequency aspect words in standard LDA. To reduce the interference of global aspect words, it is necessary to identify them as prior knowledge.

4.1. Identification of global aspect words

Global aspect words and global opinion words have obvious modification relations in the sentences, and the global aspects can be identified based on the high-frequency co-occurrence between them. We first create a set $S_{go}(w)$ containing a small number of global opinion seed words. Then, we find aspects satisfying co-occurrence frequency threshold ζ_3 from the candidate set S_{ca1}, and add them to the global aspect words $S_{ga}(w)$. We can also find the opinions corresponding to the aspects satisfying threshold ζ_3 from S_{ca1} and add them to the global opinion words $S_{go}(w)$. The abovementioned two steps continue iteratively until no new global aspects are found, and ultimately form the global aspect words. The iterative discovery process is demonstrated in Fig. 5.

4.2. Constraints in AC-LDA

In this section, we discuss how we add the constraint on global aspect words and the three types of association combinations to the LDA to conduct the probability allocation of the words to topics.

The constraint on global aspect words. When deciding the allocation probability of a global aspect w to a topic, we consider the number of existing global aspects in the topic as a key feature affecting w's topic allocation. For example, suppose topics t_i and $t_j$include n and m global aspects respectively. If $n > m$, then the weight of the word w assigned to t_i is set to be larger than that assigned to t_j. In this way, the global aspects would be assigned to a small number of topics.

The constraint on association combinations. If $w$is not a global aspect word, we first find its opinion word w_p in the sentence. Then, we determine whether (w_p, w) exists in S_{NA}, S_{LO} and S_{SA}. If yes, we add the constraint on association combinations to the LDA model. Finally, we adjust the allocation probability of w_p and w to a topic according to the associative strength of (w_p, w). That is, w_p and w will have a higher probability to be allocated to the same topic comparing to the original LDA.

4.3. AC-LDA design

AC-LDA introduces constraint knowledge to guide word-topic allocation and retains the topic words clustering function of the LDA model. The AC-LDA model is shown in Fig. 6, and important notations are summarized in Table 1.

In Fig. 6, on the one hand, we divide the topics into the global topics z_{gl} and the local topics z_{loc} based on the document-topics distribution θ of the standard LDA, the purpose is to gather global aspect words, reduce their influences on the local words distribution, and enhance the ability of discovering low-frequency local aspect words. On the other hand, we modify the topic-words distribution φ of the standard LDA model.

The document generation process of AC-LDA is demonstrated in Algorithm 1.

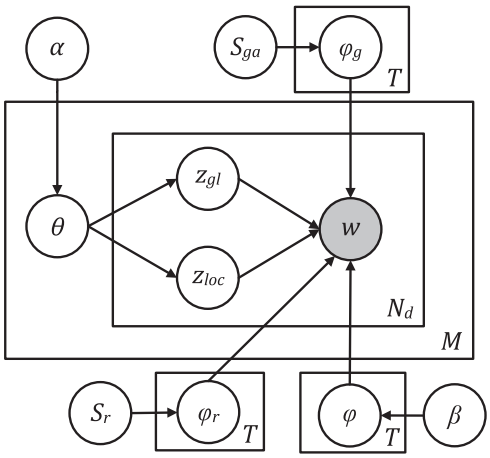


Fig. 6. An AC-LDA model.

Table 1
Notations in AC-LDA.

Notation	Description	Notation	Description
α	Dirichlet parameter of document-topics distribution	T	Number of topics
β	Dirichlet parameter of topic-words distribution	M	Number of documents
θ	Document-topics distribution	N	Number of words in a document
φ	Topic-words distribution	S_{ga}	Global aspect words
w	Word	φ_g	Global topic-words distribution
z_{gl}	Global topic	S_r	Association combinations
z_{loc}	Local topic	φ_r	Constrained topic-words distribution

Algorithm 1 Document Generation Process Algorithm of AC-LDA.

draw topics distribution θ~*Dirichlet*(α)
/* choose words distribution: */
if *words*$\in S_{ga}$ **then** choose words distribution φ_g~ζ^g*Dirichlet*(β)
/* here, ζ^g is the weight factor of global aspect words */
else choose words distribution φ~*Dirichlet*(β)
end if
for ***each*** word w_i ***in*** sentence s_k of document d, **do**
choose topic z_j~θ
if w_i is a global aspect word, **then** generate word w_i~φ_g
else if (w_p, w_i)$\in S_r$ **then** choose word w_i~φ_r /* i. e., w_p is the word adjacent before w_i */
else generate word w_i~φ
end if
end for

4.4. Inference with AC-LDA

In AC-LDA, the parameters θ and φ need to be inferred. To infer the two parameters, we must work out the probability equation of Gibbs sampling. The AC-LDA model is based on the standard LDA but incorporated with constraint variables. The Gibbs sampling probability equation can be written as follows.

$$P(z_i = k|\boldsymbol{w}, \boldsymbol{z}_{-i}, \alpha, \beta, \gamma) = \frac{P(\boldsymbol{w}, \boldsymbol{z}, \alpha, \beta, \gamma)}{P(\boldsymbol{w}, \boldsymbol{z}_{-i}, \alpha, \beta, \gamma)} \propto \frac{P(\boldsymbol{w}, \boldsymbol{z}, \alpha, \beta, \gamma)}{P(\boldsymbol{w}_{-i}, \boldsymbol{z}_{-i}, \alpha, \beta, \gamma)}, \tag{5}$$

$$\frac{P(\boldsymbol{w}, \boldsymbol{z}, \alpha, \beta, \gamma)}{P(\boldsymbol{w}_{-i}, \boldsymbol{z}_{-i}, \alpha, \beta, \gamma)} = \frac{P(\boldsymbol{w}, \boldsymbol{z}|\alpha, \beta, \gamma)}{P(\boldsymbol{w}_{-i}, \boldsymbol{z}_{-i}|\alpha, \beta, \gamma)}. \tag{6}$$

Combining Eqs. (5) and (6), we get,

$$P(z_i = k|\boldsymbol{w}, \boldsymbol{z}_{-i}, \alpha, \beta, \gamma) \propto \frac{P(\boldsymbol{w}, \boldsymbol{z}|\alpha, \beta, \gamma)}{P(\boldsymbol{w}_{-i}, \boldsymbol{z}_{-i}|\alpha, \beta, \gamma)} = P(w_i, z_i = k|\alpha, \beta, \gamma). \tag{7}$$

Expanding Eq. (7), we have,

$$P(w_i, z_i = k|\alpha, \beta, \gamma) = P(w_i|z_i = k, \alpha, \beta, \gamma)P(z_i = k|\alpha, \beta, \gamma) = P(w_i|z_i = k, \beta, \gamma)P(z_i = k|\alpha)$$

$$= P(w_i|z_i = k, \beta)P(w_i|z_i = k, \gamma)P(z_i = k|\alpha). \tag{8}$$

Based on Eqs. (7) and (8), we get,

$$P(z_i = k|\boldsymbol{w}, \boldsymbol{z}_{-i}, \alpha, \beta, \gamma) \propto P(w_i|z_i = k, \beta)P(w_i|z_i = k, \gamma)P(z_i = k|\alpha), \tag{9}$$

where

$$P(w_i|z_i = k, \beta) = \frac{\prod_k \frac{\prod_i \Gamma(N_{w_i,k}+\beta)}{\Gamma(N_k+V\beta)}}{\left(\prod_k \frac{\prod_i \Gamma(N_{w_i,k}+\beta)}{\Gamma(N_k+V\beta)}\right)_{-i}} \propto \frac{\{N_{w_i,k}\}_{-i}+\beta}{\{N_k\}_{-i}+V\beta}, \tag{10}$$

$$P(w_i|z_i = k, \gamma) = \begin{cases} 1 + R(w_p, w_i) & < w_p, w_i > \in S_r \wedge w_p \in z_k \\ 1 + \frac{1}{n_k^{gl}} \sum_{j=1}^{n_k^{gl}} P_k(w_j^{gl}) & w_i \in S_{ga}, \end{cases} \tag{11}$$

and

$$P(z_i = k|\alpha) = \frac{\prod_d \frac{\prod_k \Gamma(N_{d,k}+\alpha)}{\Gamma(N_d+T\alpha)}}{\left(\prod_d \frac{\prod_k \Gamma(N_{d,k}+\alpha)}{\Gamma(N_d+T\alpha)}\right)_{-i}} \propto \frac{\{N_{d,k}\}_{-i}+\alpha}{\{N_d\}_{-i}+T\alpha}. \tag{12}$$

In Eqs. (10)–(12), $\{N_{w_i,k}\}_{-i}$ denotes the number of words w_i assigned to topic k in addition to this time; $\{N_k\}_{-i}$ denotes the number of all words except current w_i assigned to topic k in addition to this time; n_k^{gl} denotes the number of global aspect words assigned to topic k; $P_k(w_j^{gl})$ is the probability of global aspect word w_j^{gl} assigned to topic k; $\{N_{d,k}\}_{-i}$ denotes the number of document d containing w_i assigned to topic k in addition to this time; $\{N_d\}_{-i}$ denotes the number of document d containing w_i assigned to all topics in addition to this time; V is the number of all words, Γ is Gamma function.

Lastly, we can get the Gibbs sampling probability as:

$$P(z_i = k|\boldsymbol{w}, \boldsymbol{z}_{-i}, \alpha, \beta, \gamma) \propto P(w_i|z_i = k, \gamma)\frac{\{N_{w_i,k}\}_{-i}+\beta}{\{N_k\}_{-i}+V\beta} \cdot \frac{\{N_{d,k}\}_{-i}+\alpha}{\{N_d\}_{-i}+T\alpha}. \tag{13}$$

From Eq. (13) we can get the equation calculating θ and φ, as follows.

$$\theta_{k,d} = \frac{\{N_{d,k}\}_{-i}+\alpha}{\{N_d\}_{-i}+T\alpha}, \tag{14}$$

$$\varphi_{w_i,k} = P(w_i|z_i = k, \gamma)\frac{\{N_{w_i,k}\}_{-i}+\beta}{\{N_k\}_{-i}+V\beta}. \tag{15}$$

4.5. Complexity analysis

Now we discuss the computational complexity of our AC-LDA along with the LDA and ASUM (aspect and sentiment unification model), two benchmarks we will empirically compare.

The LDA model follows a three-layer structure, including document, topic, and word. The computational complexity is approximately $O(M \times T \times N)$, where M is the number of documents; T is the number of topics; N is the number of words; and $M \gg N \gg T$. By assuming that each sentence corresponds to an aspect, the ASUM model contains four layers, including document, topic, sentence, and word. Due to the introduction of the prior knowledge of sentiment seed words, the computational complexity of the ASUM can be estimated as $O(M \times T \times L \times N \times K)$, where L is the number of sentences; K is the size of the sentiment seed words; and $M \gg N \gg T, N, K$.

Our AC-LDA follows the original LDA and is formed as a three-layer structure model. In addition, we add the embedded knowledge of word associations. Given S the size of word association combinations, the computational complexity of AC-LDA is approximately $O(M \times T \times N \times S)$, and $M \gg N \gg T, S$. Thus, the complexity of AC-LDA is in between the LDA and the ASUM. Considering that both T and S are much less than M, the AC-LDA will only lead to a limited increase in computational cost compared to the LDA.

5. Experiments and results

5.1. Data

The dataset used in the experiments consists of reviews of digital cameras on Taobao (www.taobao.com). Regarding the data, we follow [3,5,7–11,14,15,18–21,25] that focus on selected electronic products (e.g., digital camera) reviews in their experiments. It is widely accepted in the literature on aspect and opinion extraction analysis to use digital camera reviews only. Thus, we collected extensive such texts from Taobao.com, the best online shopping platform in China. A total of 101,573

Table 2
The statistical information on dataset.

Reviews	Sentences	Words per review	Vocabulary size	Nouns & gerunds	Adjectives
45,928	325,427	69.14	12,316	3,695	1,863

Table 3
Important statistics of the labeling results.

Class	Sample Size	Subclass	Sample size	Details	Sample size
Aspects	208	Global aspects	29		
		Local aspects	179	Ordinary	147
				Secondary	32
Opinions	186	Global opinions	14		
		Local opinions	172	Ordinary	140
				Low frequency	32
Non-aspect opinions	143	Global opinions	10		
		Local opinions	133		
Association combinations	269	LO association combinations	230		
		SA association combinations	39		

reviews were collected using the crawler software to set the keyword “digital camera”. In order to get enough information about the reviews, the reviews with less than 50 words were deleted, resulting in 45,928 reviews as the final full sample. We used ICTCLAS [27] to segment the Chinese sentences and applied LTP [28] to conduct dependency parsing. In our experiments, words with part of speech as nouns, gerunds, and adjectives are retained. Important statistical information regarding the final sample is reported in Table 2.

We manually labeled the aspect words, the opinion words, and the association combinations as the ground truth. Important statistics of the labeling results are summarized in Table 3. The data is labeled with multiple sets of manual annotations, combined with the evaluation description provided by Taobao. The aspect words, opinion words, non-aspect opinion words and association combinations are labeled based on the broadest consensus.

5.2. Evaluation metrics and benchmarks

We use precision, recall, and Rand index to evaluate the models. The precision measure is defined as:

$$P_a = \frac{\sum_{k=1}^{T} N_k^a}{\sum_{k=1}^{T} N_k^a @Top{-}n},\ P_o = \frac{\sum_{k=1}^{T} N_k^o}{\sum_{k=1}^{T} N_k^o @Top{-}n},\ P_c = \frac{\sum_{k=1}^{T} N_k^c}{T \cdot Top{-}n/2},$$

where P_a, P_o, and P_c denote the precision of aspect words, opinion words and association combinations, respectively; N_k^a, N_k^o and N_k^c denote the number of aspect words, opinion words and association combinations extracted accurately in topic k, respectively; $N_k^a @Top{-}n$ denotes the *top-n* noun and verb-noun extracted in topic k, and $N_k^o @Top{-}n$ denotes the *top-n* adjective extracted in topic k; T denotes the number of topics.

Recall measure is defined as:

$$R_a = \frac{\overline{N}^a @Top{-}n}{\overline{N}^a},\ R_o = \frac{\overline{N}^o @Top{-}n}{\overline{N}^o},\ R_c = \frac{\overline{N}^c @Top{-}n}{\overline{N}^c},$$

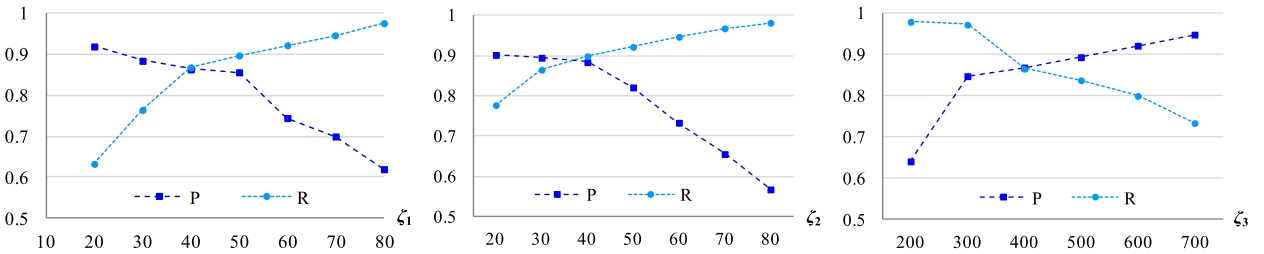
where R_a, R_o and R_c denote the recall of aspect words, opinion words, and association combinations, respectively; $\overline{N}^a @Top{-}n$, $\overline{N}^o @Top{-}n$ and $\overline{N}^c @Top{-}n$ denote the number of aspect words, opinion words, and association combinations extracted accurately without repetition in *top-n* words of all topics, respectively; $\overline{N}^a$, $\overline{N}^o$ and $\overline{N}^c$ denote the number of aspect words, opinion words, and association combinations in manually-labeled set, respectively.

A rand index is defined as:

$$RI = \frac{TP + TN}{TP + FP + FN + TN},$$

where TP denotes the number of two words that are in the same association combination, and allocated to the same topic; TN denotes the number of two words with different association combinations and allocated to different topics; FP denotes the number of two words having different association combinations and allocated to the same topic; FN denotes the number of two words that are in the same association combination and allocated to different topics.

Considering that the strength of the ASUM (aspect and sentiment unification model) in the joint extraction of aspects and opinions has already been shown in [3]. It is considered as the state-of-the-art technology in the joint extraction task so that we benchmark our AC-LDA model against the ASUM as well as a standard LDA.



(a) Precision and recall with ζ_1 (b) Precision and recall with ζ_2. (c) Precision and recall with ζ_3.

Fig. 7. Results under Different Thresholds.

5.3. Experimental settings

The parameters of topic models are estimated by the Gibbs sampling method. The ratio of the size of the test set vs. the size of the training set is 1:10. Other parameters are set as: The document-topic probability distribution parameter α is $50/T$, T is the number of topics, and the number of topic words is 20, e.g., *top-n* $= 20$; the topic-word probability distribution parameter β is 0.01; the number of sampling times is 1000, using 10-fold cross-validation.

The weight factor ζ^g of global aspect words is calculated as follows:

$$\zeta^g(w, z_i = k) = 1 + \frac{\zeta(w, z_i = k) - min_\zeta}{max_\zeta - min_\zeta},$$

where $\zeta(w, z_i = k) = \lambda n_k^{gl} - (1 - \lambda) n_k^{ng}$; n_k^{gl} is the number of global aspect words in the topic k, n_k^{ng} is the number of non-global aspect words in the topic k, and the flexible factor $\lambda = 0.6$.

The weight factor ζ^r of association combinations is computed as:

$$\zeta^r(w, z_i = k) = 1 + R(w, w_k^p),$$

where w_k^p is the prefix word of w under the topic k in the same sentence, and is also associated with w in S_{da}, S_{oa} or S_{sa}, $R(w, w_k^p)$ is the associative strength of w and w_k^p.

ζ_1, ζ_2, and ζ_3 denote the threshold to obtain LO associations, to discover SA associations, and to identify global aspect words, respectively. The results of precision and recall are shown in Fig. 7. The thresholds are set to $\zeta_1 = 40$, $\zeta_2 = 38$ and $\zeta_3 = 400$, respectively. When ζ_3 =400, through the iteration process as shown in Fig. 5, we can get 29 global aspect words (including 25 words belonging to the ground truth) and 17 global opinion words (including 13 words belonging to the ground truth).

For the ASUM, following [3], we set $\alpha = 0.1$. For positive opinion-aspect, we set $\beta = 0$ for the negative seed words and let $\beta = 0.001$ for others. Similarly, for negative opinion-aspect, we set $\beta = 0$ for the positive seed words and let $\beta = 0.001$ for other words. The number of topic words is 20; the number of sampling times is 1000. All results are reported based on 10-fold cross-validation.

5.4. Result analysis

In this section, the proposed AC-LDA model is tested and compared with the ASUM model [3] as well as the standard LDA model. The overall extraction performance of aspect words, opinion words, and word relevance (i.e. topic clustering effect) are analyzed when the constraints on global aspects and association combinations (*NA associations, LO associations, SA associations*) are added to the LDA model. The comparative analysis is done from five aspects: "extracting aspect and opinion words", "extracting aspect words", "extracting opinion words", "extracting association combinations" and "rand index analysis". As a preview of the experimental results, we show that the proposed AC-LDA model outperforms all tested benchmarks for the extraction of aspect words, opinion words, and word relevance, especially in the extraction of secondary aspect words, low-frequency opinion words and non-aspect opinion words. Moreover, our model has a better aspect clustering ability, indicating that the constraints on global aspects and association combinations can improve the semantic understanding ability of the LDA model.

Extracting aspect and opinion words. The comparison of LDA, ASUM, and AC-LDA is reported in Table 4, and the topic number T is 80. We select three topics {"appearance", "screen" and "price"} as examples here. From Table 4, it can be found that in the LDA model, the number of global aspects extracted (e.g. "camera", "function", "feeling", "logistics" and "goods", etc.) is more than other models. Regarding the ASUM model, because it takes into account the local structure of the sentence, the frequency of global words in each topic is lower than that of the LDA model.

To sum up, our AC-LDA model leads to the following good reuslts.

- Due to the introduction of the constraint on global aspect words, AC-LDA can reduce the global aspects interference, and extract more local aspects, *e.g.*, "appearance", "screen" and "price".

Table 4
Top-n topic words (translated from Chinese) from different models.

Topic	Model	*Top*-20 phrases
Topic "appearance"	LDA	Appearance, camera, clear, not bad, function, feeling, sorry, texture, touch, full-functional, thoughtful, family, color, effect, logistics, photo, picture, goods, dark, evaluation
	ASUM	Good, appearance, workmanship, praise, pixel, beautiful, cost-performance, compact, clear, camera, patience, many, feeling, fine, convenient, carry, grand, exquisite, low grade, detail
	AC-LDA	Appearance, perfect, elegant, exquisite, liberality, good look, shape, fashionable, beautiful, worn, grand, domineering, quality, reliable, outside, light, weight, surface, shutter, picture
Topic "screen"	LDA	Screen, snowflakes, high, resolution, pixel, feeling, price, dark, liquid crystal, difference, body, requirement, computer, living, thin, trouble, camera, blurred, model, light
	ASUM	Screen, slow, pixel, resolution, compact, good, low, appearance, body, defect, reaction, touch screen, price, many, logistics, computer, trouble, disappoint, sense, entirety
	AC-LDA	Screen, sensitivity, snowflake, texture, style, beautiful, pixel, high, resolution, clear, touch screen, display screen, logistics, color, photo, defect, brightness, perspective, elegant, thin
Topic "price"	LDA	Price, cheap, store, affordable, good, rapidly, better, things, low, suggestion, entity, performance, work, trace, market, guarantee, sales volume, mind, anti-counterfeiting, exquisite
	ASUM	Price, pixel, compact, affordable, appearance, delicacy, function, cheap, salable product, focal-length, graphics quality, not bad, expensive, convenience, clear, praise, glad, complete, speed, quick
	AC-LDA	Cheap, price, expensive, affordable, discount, reasonable, fair, cost, business, fame, easy, buffer, attitude, honest, character, gift, honest, general, free gift, starting

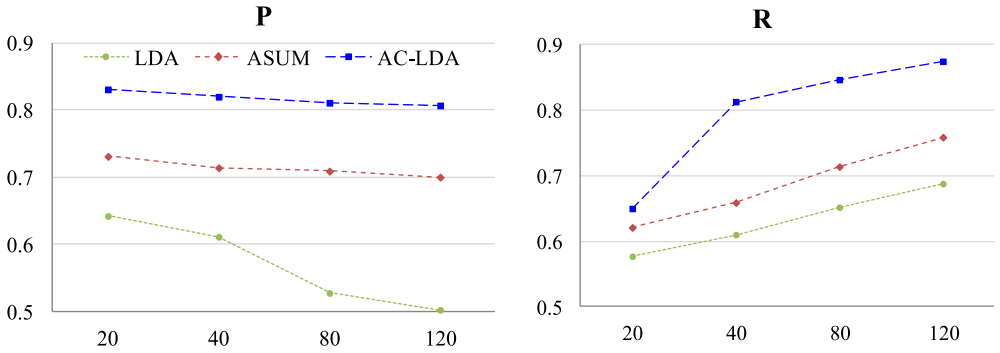


Fig. 8. Precision (P) and recall (R) of the extracted aspect words.

- There are obvious advantages in the extraction of secondary aspect words, *e.g.*, "sensitivity", "snowflake" and "brightness" related to "screen". ASUM and classic LDA cannot obtain these words.
- Although global opinion words do not appear in our AC-LDA, some low-frequency opinion words are extracted, such as "fashionable", "fair" and "reasonable", etc.
- LDA extracts some inappropriate words, such as "family", "computer", "living" and "suggestion", etc.; ASUM also extracts the word "computer". Although AC-LDA extracts the wrong word "starting", it correlates with the local aspect word "price".

Extracting aspect words. The precision (P) and recall (R) of extraction aspect words are shown in Fig. 8, where the horizontal axis T is the number of topics, and the vertical axis is the value of precision or recall. Also, we can see that AC-LDA has higher precision and recall than ASUM and LDA under all tested numbers of topics. The results can be explained as follows. When T is greater than or equal to 80, it is difficult to capture low-frequency aspect words by LDA, so the precision is lower than ASUM and AC-LDA; AC-LDA is augmented with the constraints on global aspect words, thus can reduce the interference of global aspects, and the precision does not have a significant reduction with the increase of T.

We can see that AC-LDA has clear advantages when T is high, indicating that the model is good at recognizing low-frequency aspects, while other models cannot identify these words. AC-LDA with the constraints on SA association combinations can find the secondary aspect words, such as the aspects "sensitivity" and "brightness" etc., which are associated with local aspect word "screen". In order to further compare the aspects extraction of the three models, the number of secondary aspect words extracted by each model is reported in Table 5. We can see that the AC-LDA model has a better performance on the extraction of the secondary feature words, showing that the constraints on SA association combinations and global aspect words have infected the distribution of the words.

Extracting opinion words. The precision and recall of extraction opinion words are demonstrated in Fig. 9. As can be seen, AC-LDA outperforms ASUM and LDA with the increase of T. We provide the following explanation on the results. The distribution probabilities in a topic model of low-frequency opinion words can be improved by considering the constraints on association combinations. It can be seen that the recall of AC-LDA increases obviously with the increase of T, which shows that some low-frequency opinion words are matched to the corresponding higher frequency aspect words, and these

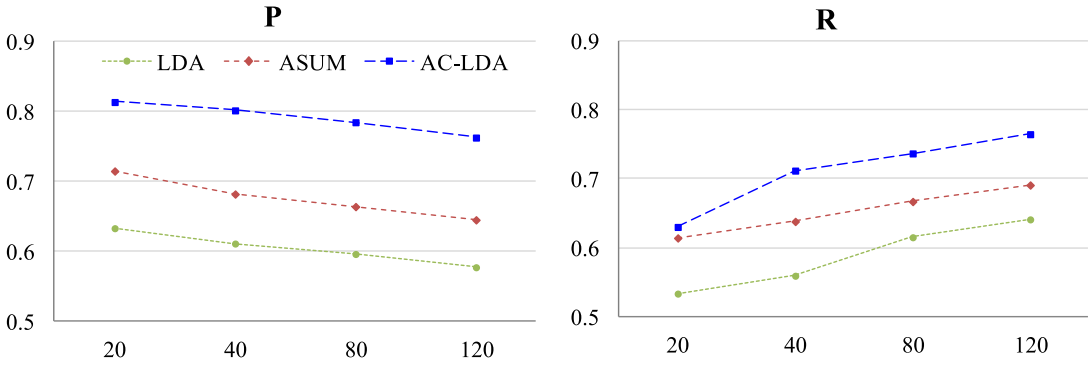


Fig. 9. Precision (P) and recall (R) of the extracted opinion words.

Table 5
The number of secondary aspect words extracted by the three models.

Model	*T* (number of topics)			
	20	40	80	120
LDA	3	7	11	12
ASUM	7	10	14	16
AC-LDA	13	19	24	26

Table 6
The number of low-frequency opinion words extracted by the three models.

Model	*T* (number of topics)			
	20	40	80	120
LDA	4	9	15	17
ASUM	9	15	18	22
AC-LDA	15	21	24	30

Table 7
The number of non-aspect opinion words extracted by the three models.

Model	*T* (number of topics)			
	20	40	80	120
LDA	68	79	87	91
ASUM	77	86	93	101
AC-LDA	85	98	112	120

opinion words can be extracted according to the aspect words. For example, some low-frequency opinion words, such as “fashionable”, “domineering”, “honest”, “fair”, etc., are not found in ASUM and LDA but found in the AC-LDA model.

The number of low-frequency opinion words extracted by the three models is shown in Table 6. The AC-LDA model has a better performance on the extraction of the low-frequency opinion words, which shows that the constraints on LO association combinations {(*local aspect word, low-frequency opinion word*)} and global aspect words infect the distribution of the words. The numbers of non-aspect opinion words extracted by the three models are reported in Table 7. As can be seen, the AC-LDA model is more effective in the extraction of the non-aspect opinion words, which indicates that the constraints of NA association combinations {(*default aspect word, non-aspect opinion word*)} and global aspect words infect the word distribution.

Extracting association combinations. The precision and recall of extraction association combinations are shown in Fig. 10, which indicates that AC-LDA has higher precision and recall than ASUM and LDA with the increase of *T*. The difference in precision of the three models is not significant when *T* is small; the reason may be that the words allocated to topics are almost high-frequency ones, and these words can be identified easily. However, with the increase of *T*, the LDA and ASUM do consider the constraints on association combinations in low-frequency words, making them difficult to find these words correctly.

LDA tends to find the high-frequency words, which leads to the words with higher distribution probability repeat more times under different topics, making the recall low. Although ASUM is augmented with the sentiment level and seed sentiment words as a priori, the assumption of one sentence containing one aspect limits the associations in complex sentences.

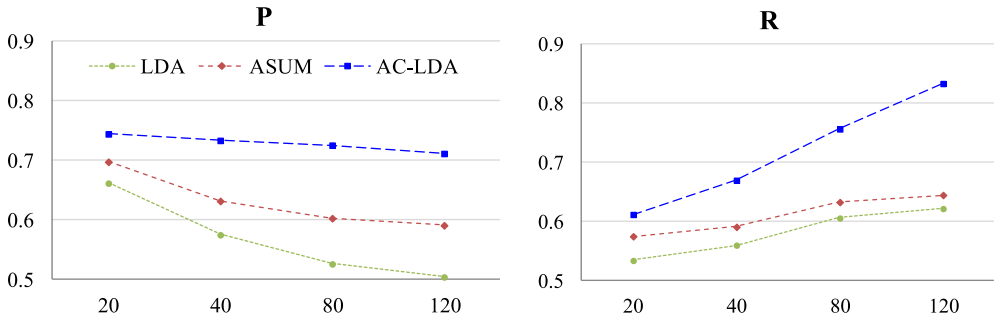


Fig. 10. The precision (P) and recall (R) of extraction association combinations.

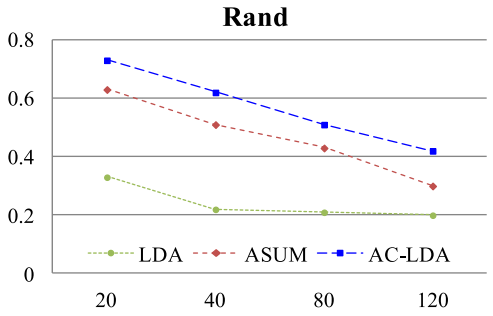


Fig. 11. Comparison of Rand Index.

The association constraints of AC-LDA make it more advantageous to find the association combinations, so the recall increases obviously with the increase of *T*. For example, the AC-LDA can assign some association combinations with relatively low co-occurrence to the same topic, such as (price, fair), etc.

Rand index analysis. The Rand index value is shown in Fig. 11. It can be seen that ASUM and AC-LDA have more obvious advantages than LDA when *T* is small, because LDA is not good at recognizing the local fine-grained association of words. The complexity and difficulty of clustering increase when *T* increases, and the overall rand value decreases. Since the association constraints are added to AC-LDA, the aspect and opinion words with low co-occurrence but a high associative degree can be classified into the same topic, so the overall rand value is better than the other two models.

5.5. AC-LDA performance analysis

The AC-LDA model intervenes in the word allocation of LDA from the perspectives of the constraints on global aspects and association combinations (*NA associations, LO associations, SA associations*). Different interventions affect the performance of the model differently. This part of the experiment compared the extent and intensity of the impact of various interventions. When the constraints on global aspect words, non-aspect opinion words, the secondary aspect words, and the low-frequency opinion words are ignored, the comparative analysis of the model's extraction ability can be better examined under different constraints. The experimental results show that for the accuracy and recall of aspect word extraction, the constraint on global aspects has the greatest influence, and the influence of the three association constraints is not very different. For the opinion words extraction, the constraints on NA associations and global aspects have higher influence on accuracy, but the most important impact on the recall rate is the constraint on LO associations, followed by the constraint on NA associations; for the association combination extraction, the impact of the four constraints on the accuracy is not very different, but the impact on the recall is dominated by the constraints on LO associations and NA associations.

The AC-LDA model attempts to improve the ability to discover the fine-grained aspect and opinion words. In order to analyze the performance of the model, we compare variants of the AC-LDA model, which incorporate the effects of different constraints as follows.

- excl-GA: Excluding the constraint of global aspect words from AC-LDA;
- excl-NA: Excluding the NA association constraint from AC-LDA;
- excl-SA: Excluding the SA association constraint from AC-LDA;
- excl-LO: Excluding the LO association constraint from AC-LDA.

Figs. 12–14. show the precision and recall of the above models. As shown in Fig. 12, the lack of prior knowledge about global aspect words has a great impact on the LDA model, and the precision and recall are significantly lower than other

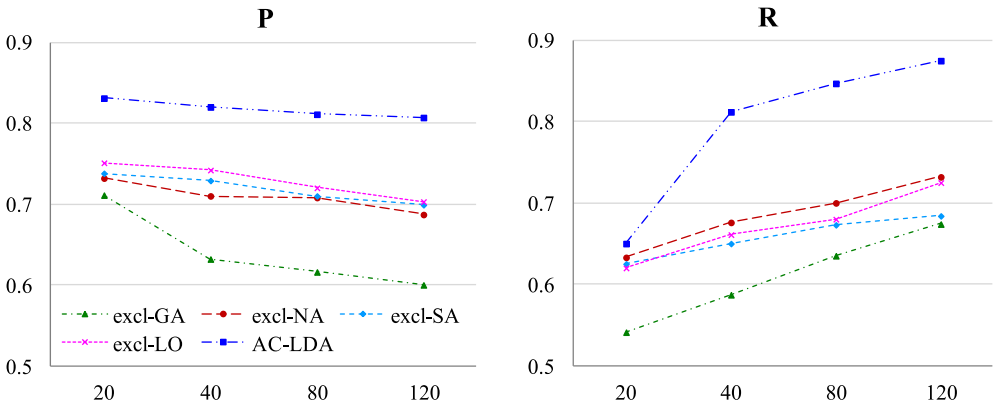


Fig. 12. The precision (P) and recall (R) of extraction aspect words.

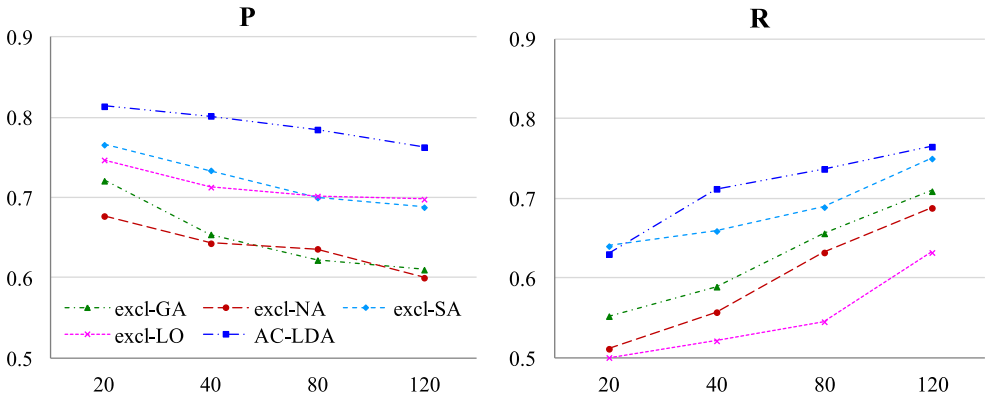


Fig. 13. The precision (P) and recall (R) of extraction opinion words.

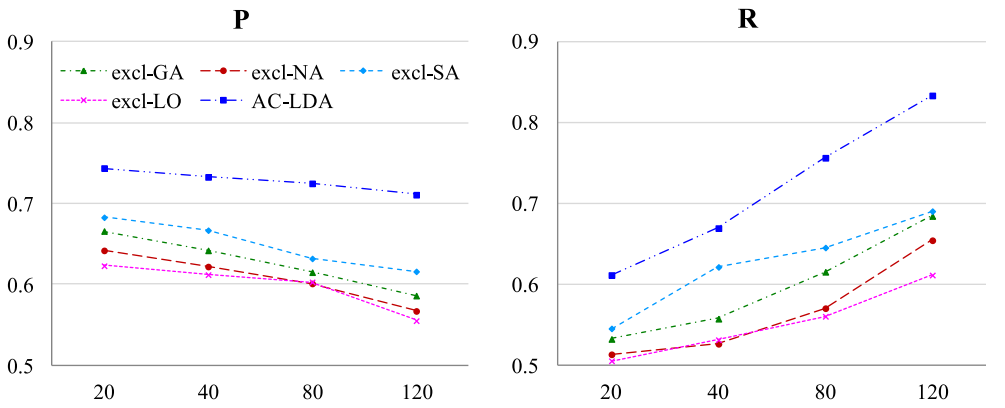


Fig. 14. The precision (P) and recall (R) of extraction association combinations.

models. Because the global aspect words are distributed almost in each topic in NG-AC-LDA, and the topic distribution probability of other aspect words is reduced, the low-frequency aspect words are difficult to be identified. In Fig. 13, we see that when there is no prior knowledge about NA association combinations in ND-AC-LDA, which reduces the ability to discover the opinion words, the precision is lower than other models. Because of the lack of prior knowledge about LO association combinations in NO-AC-LDA, other high-frequency opinion words are allocated repeatedly in different topics, which affects the discovery of low-frequency opinion words. Fig. 14 demonstrates that NG-AC-LDA has lower precision and recall, because of the lack of prior knowledge about global aspect words, the extracted association combinations are mostly global aspects or opinion words, and the distribution probabilities of local words in each topic are weakened. ND-AC-LDA

and NO-AC-LDA neglect the constraint about association combinations between the aspects and opinions, which makes it difficult to find these words, so that the precision and recall is slightly lower than that of NS-AC-LDA.

In general, our AC-LDA is a scalable algorithm for the following reasons. First, our model can be considered as an extension of LDA, and it also carries the scalability of LDA when handling large-scale text data for topic extraction. Second, the word association combinations we propose in the paper are all generalized ones, which can be used in other types of review texts. Last, for non-review texts, our model can also be utilized if corresponding embedding semantic prior knowledge can be included properly.

6. Conclusion and future work

Our work addressed the challenging task to jointly extract aspects and opinions from online product reviews in Chinese, by providing the following two innovations to existing literature. On the one hand, our AC-LDA model considers three types of word association combinations as constraints, which we formalize in this paper, and transforms the associative strengths into weight factors to affect the topic-words distribution. This design ensures that the words associated with a higher degree receive a high probability to be allocated to the same topic. Because the constraints on association combinations are extracted from sentences, our model improves the detection rate of low-frequency aspect and opinion words hidden in sentences. It also increases the coherence of the topic words. On the other hand, based on the constraint on global aspect words, the allocation distribution of global aspects is affected to concentrate on a small number of topics. In this way, the constraint reduces the impact of global aspects on other words and can improve the detection ability for local words. We obtain the global aspects from the real documents in advance and then use them as prior knowledge so that the global aspects and other words can be easily distinguished. Importantly, our model is capable of extracting aspects and opinions simultaneously. Our results show that the AC-LDA has a better performance in comparison with the state-of-the-art technique in the joint extraction task for aspects and opinions as well as the standard LDA model. Specifically, using precision and recall as the primary metrics, for the aspect extraction, our model outperforms the ASUM model by 10.4% on average in precision and 10.8% in the recall. For the opinion extraction, the precision and recall of our model are 11.5% and 5.9% higher than the ASUM model on average, respectively. For the association combination extraction, our model is also confirmed to be significantly superior to the ASUM model.

In summary, our AC-LDA model appears promising in addressing the fine-grained aspect and opinion extraction task. We also show the strength of our model in discovering the aspect-aspect and aspect-opinion associations, which can further benefit the analysis of sentence structure and words relationship in complex product review texts. From the perspective of practice, the AC-LDA model can be extended to a lot of business and management problems, such as finance (e.g., stock review analysis, financial statement studies, market attention, and sentiment estimation, etc.), human resource management (e.g., employee personality analysis), product review analysis for marketing, and the like. We would also focus on improving our model to adapt it to analyzing texts from multiple domains and resources. Other types of semi-supervised and unsupervised models to enhance the association rule generation process. Moreover, deep learning techniques may be utilized beyond the results from the AC-LDA model for further textual analysis tasks, such as sentiment classification.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Changxuan Wan: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - original draft, Writing - review & editing. **Yun Peng:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Validation, Visualization, Writing - original draft, Writing - review & editing. **Keli Xiao:** Conceptualization, Formal analysis, Methodology, Supervision, Writing - original draft, Writing - review & editing. **Xiping Liu:** Methodology, Validation, Writing - review & editing. **Tengjiao Jiang:** Methodology, Funding acquisition, Writing - review & editing. **Dexi Liu:** Methodology, Funding acquisition, Writing - review & editing.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grant Nos. 61562032, 61662032, 61662027, 61762042, 61972184 and 61966017, and the Grand Natural Science Foundation of Jiangxi Province under grant No. 20152ACB20003.

References

[1] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177.

[2] C. Lin, Y. He, Joint sentiment/topic model for sentiment analysis, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, pp. 375–384.
[3] Y. Yang, C. Chen, M. Qiu, F.S. Bao, Aspect extraction from product reviews using category hierarchy information, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017, pp. 675–680.
[4] D. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (3) (2003) 993–1022.
[5] I. Titov, R.T. McDonald, Modeling online reviews with multi-grain topic models, in: Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 111–120.
[6] D. Andrzejewski, X. Zhu, M. Craven, Incorporating domain knowledge into topic modeling via Dirichlet forest priors, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 25–32.
[7] Z. Zhai, B. Liu, H. Xu, P. Jia, Constrained LDA for grouping product features in opinion mining, in: Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2011, pp. 448–459.
[8] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, R. Ghosh, Exploiting domain knowledge in aspect extraction, in: Proceedings of the Conference on Empirical Methods on Natural Language Processing, 2013, pp. 1655–1667.
[9] Z. Chen, A. Mukherjee, B. Liu, Aspect extraction with automated prior knowledge learning, in: Proceedings of Association for Computational Linguistics, 2014, pp. 347–358.
[10] A. Bagheri, M. Saraee, F.D. Jong, ADM-LDA: An aspect detection model based on topic modeling using the structure of review sentences, J. Inf. Sci. 40 (5) (2014) 621–636.
[11] W. Zhao, J. Jiang, H. Yan, X. Li, Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2010, pp. 56–65.
[12] M. Alam, W. Ryu, S. Lee, Joint multi-grain topic sentiment: modeling semantic aspects for online reviews, Inf. Sci. 339 (2016) 206–223.
[13] B. Lu, M. Ott, C. Cardie, B.K. Tsou, Multi-aspect sentiment analysis with topic models, in: Proceedings of the 11th IEEE International Conference on Data Mining Workshops, 2011, pp. 81–88.
[14] F. Li, M. Huang, X. Zhu, Sentiment analysis with global topics and local dependency, in: Proceedings of the 24th AAAI Conference on Artificial Intelligence, 2010, pp. 1371–1376.
[15] M. Dermouche, L. Kouas, J. Velcin, S. Loudcher, A joint model for topic-sentiment modeling from text, in: Proceedings of the ACM SIGAPP Symposium on Applied Computing, 2015, pp. 819–824.
[16] C. Ma, M. Wang, X. Chen, Topic and sentiment unification maximum entropy model for online review analysis, in: Proceedings of the 24th International Conference on World Wide Web Companion, 2015, pp. 649–654.
[17] Y. Jo, A.H. Oh, Aspect and sentiment unification model for online review analysis, in: Proceedings of the 4th ACM International Conference on Web Search and Data Mining, 2011, pp. 815–824.
[18] Z. Chen, B. Liu, Mining topics in documents: standing on the shoulders of big data, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 1116–1125.
[19] G. Heyrani-Nobari, T.S. Chua, User intent identification from online discussions using a joint aspect-action topic model, in: Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014, pp. 1221–1227.
[20] M. Shams, A. Baraani-Dastjerdi, Enriched LDA (ELDA): combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction, Exp. Syst. Appl. 80 (2017) 136–146.
[21] A. Mukherjee, B. Liu, Aspect extraction through semi-supervised modeling, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012, pp. 339–348.
[22] R.K. Amplayo, S. Hwang, Aspect sentiment model for Micro reviews, in: Proceedings of the 17th IEEE International Conference on Data Mining, 2017, pp. 727–732.
[23] W. Xue, T. Li, N. Rishe, Aspect identification and ratings inference for hotel reviews, World Wide Web 20 (1) (2017) 23–37.
[24] J. Tan, A. Kotov, R.P. Mohammadiani, Y. Huo, Sentence retrieval with sentiment-specific topical anchoring for review summarization, in: Proceedings of the 26th ACM Conference on Information and Knowledge Management, 2017, pp. 2323–2326.
[25] R.K. Amplayo, S. Lee, M. Song, Incorporating product description to sentiment topic models for improved aspect-based sentiment analysis, Inf. Sci. 454 (2018) 200–215.
[26] I.A. Mel'čuk, Lexical functions, in: H. Burger, D. Dobrovol'skij, P. Kühn, N. Norrick (Eds.), Phraseology-An International Handbook of Contemporary Research, W. de Gruyter, Berlin/New York, 2007, pp. 119–131.
[27] H. Zhang, ICTCLAS: institute of computing technology, Chin. Lex. Anal. Syst. (2013) http://ictclas.nlpir.org/.
[28] W. Che, Z. Li, T. Liu, LTP: a Chinese language technology platform, in: Proceedings of the 23rd ACM International Conference on Computational Linguistics, 2010, pp. 13–16.