Event-Oriented Keyphrase Extraction Based on Bi-Clustering Model*

Lin Zhao[1,2], Liangjun Zang[1](✉), Longtao Huang[1], Jizhong Han[1,2], and Songlin Hu[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
zhaolin@iie.ac.cn, zangliangjun@iie.ac.cn, huanglongtao@iie.ac.cn, hanjizhong@iie.ac.cn, husonglin@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

Abstract. Keyphrase extraction, as a basis for many natural language processing and information retrieval tasks, can help people efficiently discover their interested information from vast streams of online documents. Previous methods are mostly proposed in general purpose, where keyphrases that represent the main topics are extracted. However, such keyphrases can hardly distinguish events from massive streams of long text documents that share similar topics and contain highly redundant information. In this paper, we address the task of keyphrase extraction for event-oriented retrieval. We propose a novel bi-clustering model for clustering the documents and keyphrases simultaneously. The model consequently makes the extracted keyphrases more specific and related to the event. We conduct a series of experiments on a real-world dataset. The experimental results demonstrate the better performance of our approach than other unsupervised approaches.

Keywords: Event-Oriented · Keyphrase Extraction · Simultaneous Learning · Bi-Clustering Model · Information Retrieval.

1 Introduction

With tremendous amounts of documents on trending and breaking news generated by various Internet media providers, it becomes increasingly difficult for people to digest such a great many streaming news information. Search engines retrieve documents from large corpora based on users' input queries that specify their interested events. However, the quality of the retrieval results depends on people's capability of refining proper keyphrases about the events. Keyphrase extraction, a task of automatically extracting descriptive phrases or concepts that represent the main topics of a document, can help people achieve more representative phrases to issue their queries. However, most existing researches on keyphrase extraction did not take the retrieval performance into consideration.

* Supported by the National Key Research and Development Program of China (No. 2017YFB1010000) and the National Natural Science Foundation of China (No. 61702500).

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22750-0_16

This causes the extracted keyphrases fail to retrieve events from massive streams of long text documents that share similar topics and contain highly redundant information. To this end, this paper addresses the task of event-oriented keyphrase extraction. The goal is to automatically extract keyphrases that represent the specific events and distinguish with others.

Accurately identifying event-oriented keyphrases from documents will benefit many downstream applications such as event-oriented information retrieval, event monitoring, event tracking, news recommendation, text summarization. For example, censors look for solutions to monitor online current affairs hotspots from blogs, forums and news site for possible criminal activities and event-oriented public opinion analysis. The key to extract event-oriented keyphrases is to find more documents about the specific events and decrease the possibility of finding documents describing other events.

Generally, previous studies about automatic keyphrase extraction fall into two categories: the supervised keyphrase extraction and the unsupervised keyphrase extraction. The supervised keyphrase extraction task is usually treated as a binary classification problem [8,14]. In this approach, keyphrases and non-keyphrases are labeled by human judges in training documents and a classifier is trained by using training documents annotated with keyphrases, then the classifier is applied to determine whether a candidate phrase is a keyphrase in test documents. On the contrary, the unsupervised methods do not require labeled training data. Instead, some external statistic information are explored to identify the keyphrases [18,35]. In this paper, we mainly focus on extracting keyphrases with an unsupervised method. However, the existing unsupervised keyphrase extraction methods suffer from two drawbacks:

- **Ignore users' specific needs.** The document collection may contain several events (or several aspects of an event). For different applications, the users may be interested in different events or aspects, thus they need different keyphrases to reflect their needs. Taking the theme of earthquake from Sina news as an example (see Section 4.1), there are some earthquake events, e.g. "四川康定发生6.3级地震"(Sichuan earthquake), "尼泊尔发生8.1级强震" (Nepal earthquake)", "earthquake" is a keyphrase for the entire theme while it should not be the keyphrase when we only consider the rescue aspect in Nepal earthquake.
- **Fail to extract event-oriented keyphrases.** Considering different events with simi-lar topics are likely to share keyphrases, which makes existing methods fail to distinguish a specific event from others with similar topics. For example, topic models (e.g. Latent Dirichlet Allocation) can simultaneously cluster documents and generate representative words for each topic, but topic-based methods are likely to put the documents about "Sichuan earthquake" and those about "Nepal earthquake" into the same cluster because both events share many topical keywords like "坍塌" (collapse), "救援" (rescue) and "重建" (rebuild).

Existing work rarely focuses on automatically extracting keyphrases for a particular event. In this paper, we address the task of event keyphrase extraction, which is a specific task to obtain keyphrases for users about their interested events. Actually, the task of event keyphrase extraction comprises two sub-tasks, i.e. event identification and keyphrase extraction. Event identification aims to cluster all documents into different groups, each of which corresponds to one specific event. Keyphrase extraction aims to extract keyphrases from the documents for each event. A simple and straightforward strategy is to combine them in a pipelined way, but it suffers the problem that the errors raised by event identification will lead to poor performance of keyphrase extraction. Another more reasonable strategy is to simultaneously address the two subtasks. It provides the opportunity of mutual boosting, i.e., the two sub-tasks can potentially benefit from each other, therefore we adopt the way of simultaneous learning. For evaluating our approach, we implemented a retrieval system to evaluate the effectiveness of event-oriented keyphrases in the application of information retrieval. The major contributions of this paper are summarized as follows:

(1) We address the task of event-oriented keyphrase extraction, which can benefit many downstream applications.

(2) We take users' specific needs and event-oriented characteristics into consideration and propose the novel method based on the bi-clustering model.

(3) Extensive experiments over real data show that our method achieves better performance compared with the other unsupervised methods.

The rest of the paper is organized as follows. In Section 2, we give some background information and related works. Section 3 presents the methodology, including problem formulation, keyphrase extraction, and document retrieval. In Section 4, we describe the experiment and analysis, including experiment setting and results analysis. Section 5 presents conclusions and a discussion of further research.

2 Related Work

Automatic keyphrase extraction, event detection, and document retrieval are three lines of studies that are related to our work. Automatic keyphrase extraction selects the important and topical keywords from documents automatically [33] and its goal is to extract phrases that represent or are related to the topics discussed in the given documents [5, 10, 20, 31]. Event detection is a basic problem of the information extraction, and it is also a sub-field in information retrieval, which aims to detect real events from multiple document streams [7, 32, 38, 40]. Document retrieval is a process of matching the user's request against a collection of documents [22]. In this paper, we apply document retrieval to evaluate the performance of event-oriented keyphrases.

2.1 Keyphrase Extraction

Existing methods for keyphrase extraction can be divided into supervised and unsupervised methods [23]. Most supervised methods formalize keyphrase extraction as a binary classification problem. However, supervised methods require

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22750-0_16

training data with labeled keyphrases, which is extremely expensive and time-consuming for do-main-oriented application scenarios and hard to adapt to other domains.

Existing unsupervised methods for keyphrase extraction can be categorized into several groups. First, rank-based methods build a word co-occurrence graph from the input document and then ranking all nodes on the graph [12, 19, 34]. Such methods often suffer the problem of information loss. For example, if two words never occur together within a predefined window size, there will be no edge to connect them in the co-occurrence graph. Second, topic-based methods aim to group the candidate keyphrases in a document into topics, such that each topic is composed of all and only those candidate keyphrases that are related to that topic [15, 20]. Third, external resource based methods explore external text corpus to enhance the performance of keyphrase extraction [29, 30]. However, such methods often bring about the information overload problem, e.g. the real meanings of words in the document may be overwhelmed by a large amount of introduced external corpus. Finally, knowledge-based methods combine semantic similarity clustering with knowledge graphs to help discover hidden semantic relations in documents [6]. Such a method does not consider the relevance of news articles belong to the same event, which demonstrates efficiency only in single-document keyphrase extraction task.

2.2 Event Detection

Many tasks are put forward to develop and evaluate technologies for event detection, e.g. topic detection and tracking (TDT), automatic content extraction (ACE), and text analysis conference knowledge base population (TAC KBP). In addition, there has been a lot of research work in event detection. Neural network models have been the most successful methods for event detection. For instance, Chung et al. [1] introduce a DAG-GUR architecture that captures the syntactic and context information through a bi-directional reading of the text with dependency parse relation-ships. Feng et al. [7] combined a convolutional neural network (CNN) with a bi-directional long short-term memory (Bi-LSTM) to create a hybrid network. The hybrid network was fed to a linear model for capturing sequence and chunk information from specific contexts.

2.3 Document Retrieval

The main concern of document retrieval is to retrieve documents from a corpus based on the specific user query. Different probabilistic model and language model have been proposed over the past decade for document retrieval, such as BM25 [27], probabilistic retrieval model for semi-structured data [17], mixture of language models [25], and machine learning-based methods [24, 39]. However, early methods mentioned above are difficult to meet the needs of users. They assume bag-of-words document representation and match phrases directly in queries and documents, which suffers from the issue of lexical gap, when similar concepts are expressed using different words in queries and relevant documents. The success of deep learning has revitalized research on text matching recently. Several neural architectures have been proposed for document retrieval [2, 3, 26, 36, 37], which can be categorized into two classes according to their model architecture. One is representation-based model, such as DSSM [13], CDSSM [28], ARC-I [11] and DCNN [16], which builds a good representation for a text with deep neural network, and then conducts matching between text representations. The other is the interaction-based model, such as DeepMatch [21], ARC-II [11], DRMM [9] and K-NRM [36], which builds local interactions between two texts, and then learns hierarchical interaction patterns for matching with deep neural networks.

3 Methodology

The framework of our event-oriented keyphrase extraction is shown in Fig. 1. It consists of the following three steps: (1) problem formulation, (2) keyphrase extraction, and (3) document retrieval. We collect a lot of news documents, each document corresponds to the specific event and each event includes many documents. Then we utilize the bi-clustering model to extract keyphrases from news documents. Finally, we use the keyphrases as queries to retrieve news documents, in order to evaluate the ability of keyphrases to represent events.

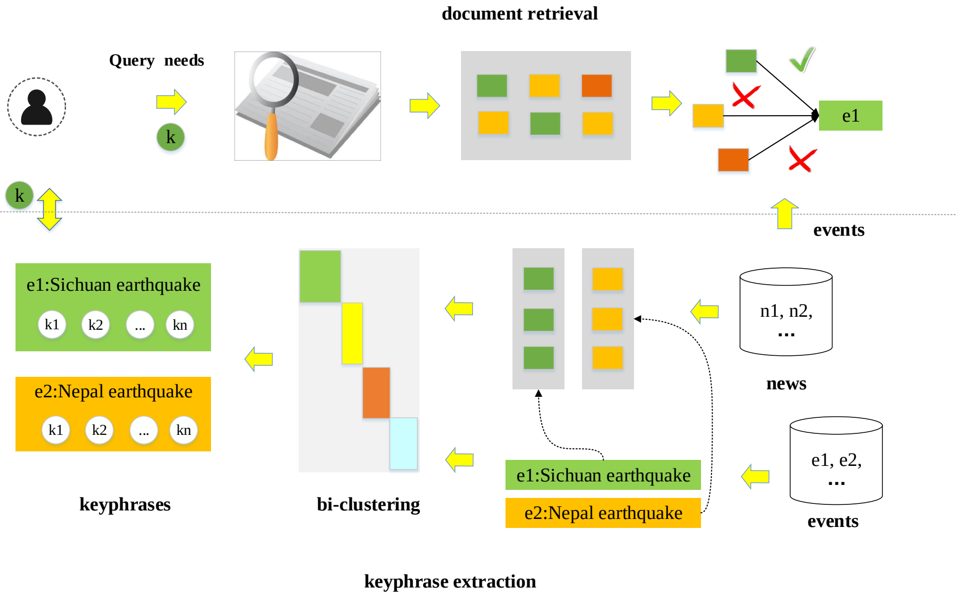


Fig. 1. The process of event-oriented keyphrase extraction

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22750-0_16

3.1 Problem Formulation

Event-Oriented Keyphrases We define event-oriented keyphrase as a continuous sequence of keywords, which is highly important and relevant to the event. For instance, “地震” (earthquake), “康定县” (Kangding County), “四川省” (Sichuan Province), and “倒塌” (collapse) are the keyphrases of the example document in Fig. 2 (a). “防御中心” (Prevention Center), “监测” (monitoring), “调查” (investigation), and “损失” (damage) are the keyphrases of the example document in Fig.2 (b). Note that “地震” (earthquake) and “倒塌” (collapse) are not the keyphrases in the event of Nepal earthquake, because 1) the event of Nepal earthquake expresses the rescue phase rather than report the earthquake in this example, and 2) they are shared in both the “Sichuan earthquake” and “Nepal earthquake”, which cannot distinguish different events with similar topics.

Keyphrases Extraction Given the event set E and news set N includes D news subset, each news subset d_j corresponds to an event e_i. Our goal is to extract event-oriented keyphrases of each event from the corresponding collection D_i, where

- $E = \{e_i | 1 \leq i \leq |E|\}$ denotes the collection of specific events.
- $D = \{d_j | 1 \leq j \leq |D|\}$ denotes the news subsets.
- $N = \sum\limits_{i=1}^{|E|} D_i$ denotes the news data set grouped by events.

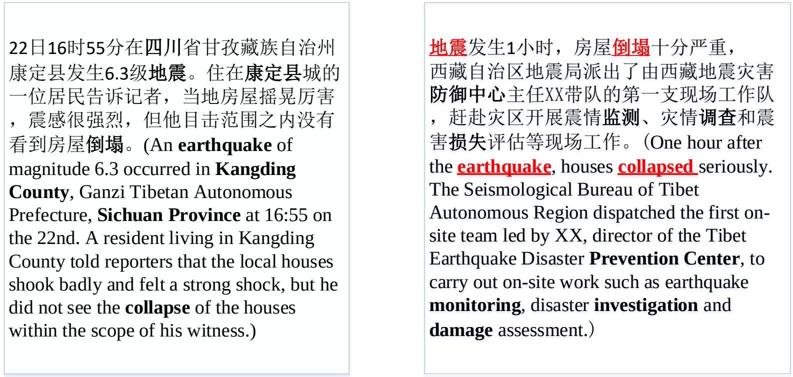
Keyphrase Evaluation We evaluate event-oriented keyphrases through event-oriented retrieval tasks. Given the event-oriented keyphrases K by keyphrase extraction mentioned above. We issues $K \in e_i$ as queries to retrieve news in search engine (e.g. Google Search). The search engine returns a collection of news set R. The goal is to evaluate the effectiveness of K by judging whether set R belongs to e_i.

- $K = \{k_1, k_2, ..., k_{|K|}\}$ denotes the collection of event-oriented keyphrases are extracted from D_i.
- $R = \{r_1, r_2, ..., r_{|R|}\}$ denotes the retrieved news collection related to the event e_i.

3.2 Keyphrases Extraction

Feature Selection Different from the keyphrase extraction in document level, we consider the correlation of keyphrases as event features in specific events and propose the following hypothesis:

Assume that an event is represented by a set of key elements. Event features are used to enrich these key elements. There are the following relationships:



22日16时55分在**四川**省甘孜藏族自治州康定县发生6.3级**地震**。住在**康定县**城的一位居民告诉记者，当地房屋摇晃厉害，震感很强烈，但他目击范围之内没有看到房屋**倒塌**。(An **earthquake** of magnitude 6.3 occurred in **Kangding County**, Ganzi Tibetan Autonomous Prefecture, **Sichuan Province** at 16:55 on the 22nd. A resident living in Kangding County told reporters that the local houses shook badly and felt a strong shock, but he did not see the **collapse** of the houses within the scope of his witness.)

(a) Sichuan earthquake

地震发生1小时，房屋**倒塌**十分严重，西藏自治区地震局派出了由西藏地震灾害**防御中心**主任XX带队的第一支现场工作队，赶赴灾区开展震情**监测**、灾情**调查**和震害**损失**评估等现场工作。（One hour after the **earthquake**, houses **collapsed** seriously. The Seismological Bureau of Tibet Autonomous Region dispatched the first on-site team led by XX, director of the Tibet Earthquake Disaster **Prevention Center**, to carry out on-site work such as earthquake **monitoring**, disaster **investigation** and **damage** assessment.）

(b) Nepal earthquake

Fig. 2. Example of event-oriented keyphrases from ”Sichuan earthquake” (a) and ”Nepal earthquake” (b). Shared keyphrases are red font and underlined in ”Nepal earthquake”.

(1) Key elements extracted from similar topics are similar. (2) Event features are to enrich a group of key elements. For instance, the three key elements of the event of earth-quake = “四川” (Sichuan province in China), “地震” (earthquake), “记者” (journalist). The features are “康定县” (Kangding country), “应急” (emergency), “刘忠俊” (Zhongjun Liu), which are to describe “四川” (Sichuan province in China), “地震” (earthquake), “记者” (journalist) respectively.

Based on the above hypothesis, the method of extracting features include two steps: (1) Find the set of key elements. First, we select candidate elements with named entity recognition from each news. Then we calculate the word frequency of each candidate element in the entire news dataset. Finally, the target key elements with maximum word frequency are selected. (2) Find corresponding features. We obtain candidate features using a stop word list to remove stop words and filtering words with certain part-of-speech tags (e.g., nouns, verbs). Then we exploit the co-occurrence relation between features and key elements. The feature is selected if this feature and an element appear simultaneously in a sentence.

Bi-Clustering Modeling Bi-clustering consists of simultaneous partitioning of the set of samples and the set of their attributes (features) into subsets (classes). Samples and features classified together are supposed to have high relevance to each other [6]. Motivated by this work, we employ the bi-clustering process to analyze the relationship between news documents and event features. A basic premise behind our method is a duality of features and document clustering. That is, feature clustering induces document clustering while document clustering induces feature clustering. The goal is to find the best bi-clusters with correlation higher than those in the corresponding other rows and columns.

Generally, given news and features to bi-clustering, the process can be summarized as follows:

(1) Each of the news documents n to be clustered is represented by a p-dimensional feature vector. The entire documents are represented by a matrix of shape $n \times p$. We calculate TF-IDF values of each of the features, and a word frequency matrix (denoted as $A_{n\times p}$) is obtained by values vectorizing. We apply log normalization to normalize the matrix $A_{n\times p}$. The log of the data matrix is computed with $L = logA_{n\times p}$. The final matrix is computed according to the formula:

$$K_{ij} = L_{ij} - \overline{L_{i\cdot}} - \overline{L_{\cdot j}} + \overline{L_{\cdot\cdot}} \tag{1}$$

where $\overline{L_{i\cdot}}$ denotes the column mean, $\overline{L_{\cdot j}}$ denotes the row mean and $\overline{L_{\cdot\cdot}}$ denotes the overall mean of L.

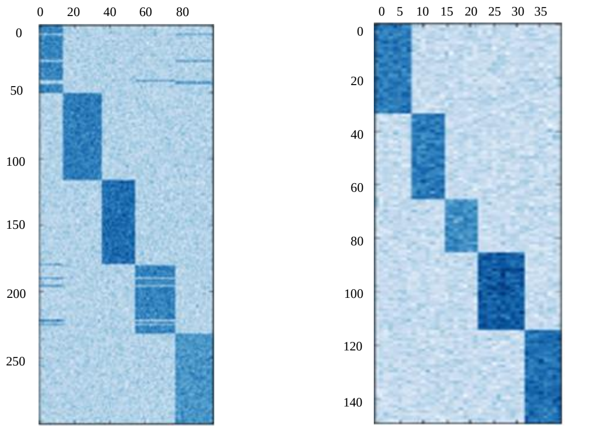
(2) We take a matrix K as input to bi-clustering and get the bi-cluster partitions with Dhillon's Spectral Co-Clustering algorithm [4]. Ideally, the rearranging the matrix reveals the bi-clusters on the diagonal because each row and each column belongs to exactly on bi-cluster. But there are some noise data in actual news, so the diagonal structure is not perfect.

(3) We filter out the noise data from sub-matrix in the diagonal as the best features bi-clusters. Each of the bi-clusters corresponds to an event. The best bi-clusters indicate subsets features used more often in those subsets news. For instance, the bi-clustering results of raw data and filtering data are shown in Fig. 3 (a) and (b), respectively, where the raw data come from our experimental dataset. First, we select 100 news documents and corresponding 300 features, Fig. 3 (a) presents the bi-clustering result. There are five bi-cluster partitions in Fig. 3 (a), where the first and the fourth bi-clusters contain noise data. Then we filter out the noise data from the two bi-clusters and run the bi-clustering process again, the result is shown in Fig. 3 (b).

3.3 Document Retrieval

Our task of event-oriented keyphrases extraction is to retrieve events from massive streams of long text documents. Hence, we can issue event-oriented keyphrases as queries to the document retrieval system, and evaluate the quality of event-oriented keyphrases by returning the results of the retrieved document (see Fig. 4).

It is ambiguous for using raw keyphrase to search event directly in a corpus containing lots of similar topic collections because it may appear in multiple events. For instance, we use “地震” (earthquake) as the query to search in a search engine, the returned retrieval lists will be much news about the different event. Hence, preparing for information retrieval tasks, we need to generate query suggestions based on the extracted keyphrases. The method of generating suggestion is the combination of keyphrases. Keyphrases combination can narrow down the scope of the search and help users to find more relevant news. The detailed processes are as follows:



(a) bi-clustering with raw data

(b) bi-clustering with filtering data

Fig. 3. An example of bi-clustering before and after filtering the raw data

(1) Query suggestion generation. Each query corresponds to a user's need for event queries. Query suggestions are generated based on different events. We apply a query parser to map the relationship between the query and the event. The mapping method is to calculate the news numbers of different events and select the event with maximum news numbers as the target event.

(2) We issue different queries (e.g. event title, query suggestion) into a document retrieval system based on BM25 in order to compare the quality of different queries.

(3) We calculate the accuracy, recall, and F1-measure based on the retrieved results returned by different queries. The higher the three metrics, the better the quality of the keywords.

4 Experiment and Analysis

In this section, we describe the process of experiment and evaluation result of testing our bi-clustering model on the dataset. We first describe the dataset collected by crawling from Sina news, then we introduce the evaluation tasks and metrics for the up-coming event-oriented retrieval. Finally, we compare and analyze the results from the bi-clustering model and baselines in evaluation tasks.

4.1 Experiment Setting

Data Description To our knowledge, there are no datasets available to satisfy our task currently. We picked 52 events ranged from 2014 to 2016 from Sina News, including 5932 news. The criterion we used for selecting events is that they should contain multiple categories and have similar topics in different categories. These events are involved in the categories of disaster, accident, conference,

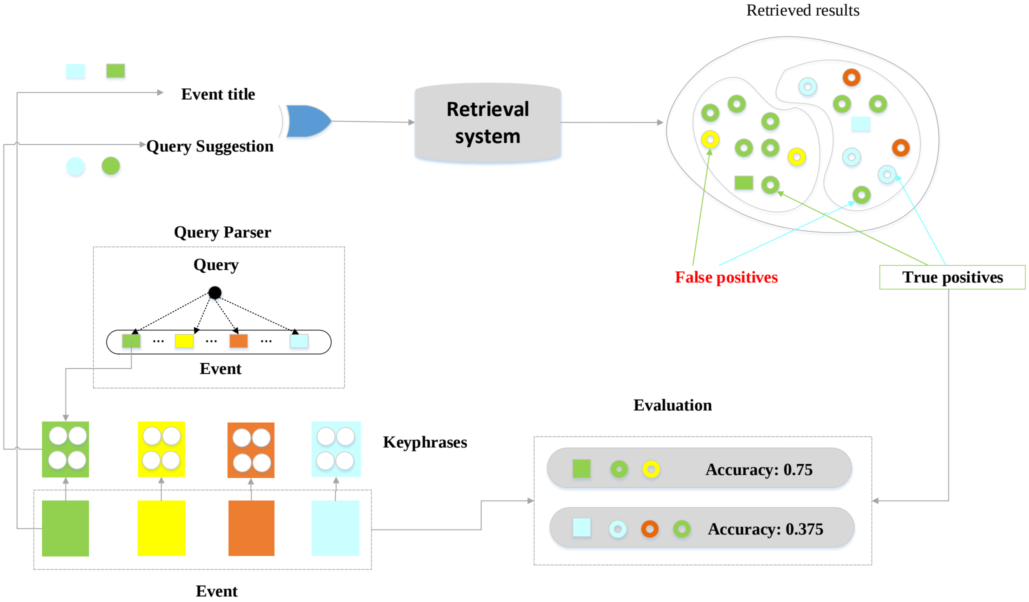


Fig. 4. The process of document retrieval based on event-oriented keyphrases

current affairs, and military. Table 1 shows some examples of the selected events and corresponding category. We hired several annotators to create truth data. Annotators were asked to (1) crawl news collection from Sina News, (2) classify news documents into events corresponding to the above-mentioned categories. Finally, we extract event-oriented keyphrases from 80% news of each category and the remaining 20% news are used for event-oriented retrieval tasks.

Evaluation Metrics For evaluating the quality of event-oriented keyphrases, we consider the following queries generated by different methods for performance comparisons: (i) event titles as queries directly, (ii) query suggestions based on the topic model as queries, and (iii) query suggestions based on the bi-clustering model. We collect the top-K news documents ($K = 10$) of the query results. The metrics for comparison are precision, recall, and F1-measure. For the returning results of each event, these metrics were defined as follows:

$$p_e = \frac{\sum_{q \in Q} p_q}{|Q|} \tag{2}$$

$$r_e = \frac{\sum_{q \in Q} r_q}{|Q|} \tag{3}$$

$$f_{measure-e} = \frac{2 \times p_e \times r_e}{p_e + r_e} \tag{4}$$

Table 1. The distribution of selected events and category.

Category	Number	Event sample
Disaster	10	四川省康定县6.3级地震 (A 6.3 magnitude earthquake in Kangding, Sichuan)
Accident	10	俄罗斯客机在埃及坠毁 (The Russian airliner crashed in Egypt)
Conference	5	全国人民代表大会常务委员会第十一次会议 (The eleventh meeting of the Standing Committee of the 12 National People's Congress)
Current affairs	19	习近平访问瑞士 (Xi Jinping's visit to Switzerland)
Military	8	中国人民解放军军改 (Military reform of the Chinese Liberation Army)

where Q is the number of query suggestions in each event, p_q is the precision of each query suggestion and r_q is the recall of each query suggestion.

4.2 Results Analysis

We present the results of precision, recall, and F1-measure based on queries generated by different methods in Table 2.

We have the following observations from Table 2: (1) From the results on all categories, bi-clustering model outperforms the other two both on precision and recall. The topic model performs better than the event titles on precision and recall. In our consideration, event titles as queries can find key elements and corresponding features in order to match target news accurately, while other methods are missing important event features. For instance, there is returned seldom results using the title "2016各地两会" (The National People's Congress and Chinese People's Political Consultative Conference around the nation in 2016). The reason is that "各地" (around the nation) does not appear in news, it was expressed using the name of different cities, such as "四川" (Sichuan province in China), "成都" (Chengdu city in Sichuan province). (2) From the results on each category, we can find the disaster category performs the best and conference category performs the worse than other categories. This is because the news described different events is a higher similarity in the conference, so most of the key elements are the same. For instance, "全国人民代表大会常务委员会第十一次会议" (the eleventh meeting of the Standing Committee of the 12 National People's Congress) and "全国人民代表大会常务委员会第十一次会议" (the tenth meeting of the Standing Committee of the 12 National People's Congress) are shared the same keyphrases, so it is difficult to distinguish such events.

Table 2. Performance comparisons of different methods, the highest values are in bold.

Category	Metric	Event titles	Topic model	Bi-clustering model
Overall	P	0.665	0.809	**0.854**
	R	0.443	0.538	**0.568**
	F	0.531	0.647	**0.683**
Disaster	P	0.633	0.953	**0.980**
	R	0.422	0.635	**0.652**
	F	0.507	0.762	**0.784**
Accident	P	0.825	0.890	**0.940**
	R	0.549	0.592	**0.626**
	F	0.659	0.712	**0.752**
Conference	P	0.400	0.553	**0.647**
	R	0.264	0.368	**0.430**
	F	0.318	0.442	**0.517**
Current affairs	P	0.700	0.813	**0.833**
	R	0.466	0.542	**0.555**
	F	0.559	0.651	**0.667**
Military	P	0.766	0.840	**0.873**
	R	0.516	0.556	**0.581**
	F	0.613	0.672	**0.698**

5 Conclusion

In this paper, we have implemented the task of event-oriented keyphrases and proposed a novel method for event-oriented keyphrase extraction, which uses the bi-clustering model to cluster news documents and keyphrases simultaneously. Also, we implement a simple document retrieval system based on BM25 to evaluate the quality of the extracted keyphrases. The experimental results showed that the event-oriented keyphrases greatly improve the performance of event retrieval.

Our future work may include the following directions. First, we only evaluated the performance of the proposed model by conducting experiments on a single Chinese news dataset for engineering needs in this paper. Therefore, in the future, we will try to conduct experiments on English dataset (e.g. crawl events from Wikipedia). Then, we can extract event-oriented keyphrases by deep learning model. Finally, we plan to utilize external knowledge to compensate for event information. For example, integrate knowledge graph into the task of event-oriented keyphrases extraction to improve the performance of event retrieval.

References

1. Chung, J., Çaglar Gülçehre, Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR **abs/1412.3555** (2014)
2. Dai, Z., Xiong, C., Callan, J.P., Liu, Z.: Convolutional neural networks for soft-matching n-grams in ad-hoc search. In: WSDM (2018)

3. Dehghani, M., Zamani, H., Severyn, A., Kamps, J., Croft, W.B.: Neural ranking models with weak supervision. In: SIGIR (2017)
4. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: KDD (2001)
5. Ding, Z., Zhang, Q., Huang, X.: Keyphrase extraction from online news using binary integer programming. In: IJCNLP (2011)
6. Farzindar, A., Khreich, W.: A survey of techniques for event detection in twitter. Computational Intelligence **31**, 132–164 (2015)
7. Feng, X., Huang, L., Tang, D., Ji, H., Qin, B., Liu, T.: A language-independent neural network for event detection. Science China Information Sciences **61**, 1–12 (2016)
8. Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C., et al.: Domain-specific keyphrase extraction. In: PROC. SIXTEENTH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE. pp. 668–673. Morgan Kaufmann Publishers (1999)
9. Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A deep relevance matching model for ad-hoc retrieval. In: CIKM (2016)
10. Hasan, K.S., Ng, V.: Automatic keyphrase extraction: A survey of the state of the art. In: In Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL (2014)
11. Hu, B., Lu, Z., Li, H., Chen, Q.: Convolutional neural network architectures for matching natural language sentences. In: NIPS (2014)
12. Huang, C., Tian, Y., Zhou, Z., Ling, C.X., Huang, T.: Keyphrase extraction using semantic networks structure analysis. Sixth International Conference on Data Mining (ICDM'06) pp. 275–284 (2006)
13. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.P.: Learning deep structured semantic models for web search using clickthrough data. In: CIKM (2013)
14. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In: Proc. of the 2003 Conf. on Empirical Methods in NLP. pp. 216–223 (2003)
15. Joorabchi, A., Mahdi, A.E.: Automatic keyphrase annotation of scientific documents using wikipedia and genetic algorithms. J. Information Science **39**, 410–426 (2013)
16. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: ACL (2014)
17. Kim, J., Xue, X., Croft, W.B.: A probabilistic retrieval model for semistructured data. In: In ECIR. pp. 228–239 (2009)
18. Liu, Z., Huang, W., Zheng, Y., Sun, M.: Automatic keyphrase extraction via topic decomposition. In: EMNLP (2010)
19. Liu, Z., Huang, W., Zheng, Y., Sun, M.: Automatic keyphrase extraction via topic decomposition (2010)
20. Liu, Z., Li, P., Zheng, Y., Sun, M.: Clustering to find exemplar terms for keyphrase extraction. In: EMNLP. pp. 257–266 (2009)
21. Lu, Z., Li, H.: A deep architecture for matching short texts. In: In: Advances in Neural Information Processing Systems. pp. 1367–1375 (2013)
22. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval (2008)
23. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: EMNLP (2004)
24. Nikolaev, F., Kotov, A., Zhiltsov, N.: Parameterized fielded term dependence models for ad-hoc entity retrieval from knowledge graph. In: SIGIR (2016)

25. Ogilvie, P., Callan, J.P.: Combining document representations for known-item search. In: SIGIR (2003)
26. Onal, K.D., Altingövde, I.S., Senkul, P., de Rijke, M.: Getting started with neural models for semantic matching in web search. CoRR **abs/1611.03305** (2016)
27. Robertson, S.E., Zaragoza, H., Taylor, M.J.: Simple bm25 extension to multiple weighted fields. In: CIKM (2004)
28. Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G.: A latent semantic model with convolutional-pooling structure for information retrieval. In: CIKM (2014)
29. Shi, T., Jiao, S., Hou, J., Li, M.: Improving keyphrase extraction using wikipedia semantics. 2008 Second International Symposium on Intelligent Information Technology Application **2**, 42–46 (2008)
30. Shi, W., Zheng, W., Yu, J.X., Cheng, H., Zou, L.: Keyphrase extraction using knowledge graphs. In: APWeb/WAIM (2017)
31. Tomokiyo, T., Hurst, M.: A language model approach to keyphrase extraction. In: In Proceedings of ACL Workshop on Multiword Expressions. pp. 33–40 (2003)
32. Tu, W., Cheung, D.W.L., Mamoulis, N., Yang, M., Lu, Z.: Real-time detection and sorting of news on microblogging platforms. In: PACLIC (2015)
33. Turney, P.D.: Learning algorithms for keyphrase extraction. Information Retrieval **2**, 303–336 (2000)
34. Wan, X., Xiao, J.: Exploiting neighborhood knowledge for single document summarization and keyphrase extraction. ACM Trans. Inf. Syst. **28**, 8:1–8:34 (2010)
35. Wan, X., Yang, J., Xiao, J.: Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In: ACL (2007)
36. Xiong, C., Dai, Z., Callan, J.P., Liu, Z., Power, R.: End-to-end neural ad-hoc ranking with kernel pooling. In: SIGIR (2017)
37. Yang, L., Ai, Q., Guo, J., Croft, W.B.: anmm: Ranking short answer texts with attention-based neural matching model. In: CIKM (2016)
38. Yang, M., Cui, T., Tu, W.: Ordering-sensitive and semantic-aware topic modeling. In: AAAI (2015)
39. Zhiltsov, N., Kotov, A., Nikolaev, F.: Fielded sequential dependence model for ad-hoc entity retrieval in the web of data. In: SIGIR (2015)
40. Zhu, J., Xu, C., Li, Z., Fung, G.P.C., Lin, X., Huang, J., Huang, C.: An examination of on-line machine learning approaches for pseudo-random generated data. Cluster Computing **19**, 1309–1321 (2016)