Hindawi
International Journal of Digital Multimedia Broadcasting
Volume 2017, Article ID 8594863, 5 pages
https://doi.org/10.1155/2017/8594863

Research Article

Prior Knowledge-Based Event Network for Chinese Text

Yunyu Shi,[1] Jianfang Shan,[2] Xiang Liu,[1] and Yongxiang Xia[1]

[1]*School of Electronic & Electric Engineering, Shanghai University of Engineering Science, 333 Longteng Road, Songjiang District, Shanghai, China*
[2]*School of Information, Qilu University of Technology, 3501 Daxue Road, Changqing District, Jinan, China*

Correspondence should be addressed to Jianfang Shan; sjfshan@163.com

Received 13 January 2017; Accepted 7 May 2017; Published 4 June 2017

Academic Editor: Hyo-Jong Lee

Text representation is a basic issue of text information processing and event plays an important role in text understanding; both attract the attention of scholars. The event network conceals lexical relations in events, and its edges express logical relations between events in document. However, the events and relations are extracted from event-annotated text, which makes it hard for large-scale text automatic processing. In the paper, with expanded CEC (Chinese Event Corpus) as data source, prior knowledge of manifestation rules of event and relation as the guide, we propose an event extraction method based on knowledge-based rule of event manifestation, to achieve automatic building and improve text processing performance of event network.

1. Introduction

Text representation is an important issue in natural language processing, such as information retrieval and text classification. An appropriate representation not only can reflect text semantic, theme, and structure but also can improve the computational efficiency. In recent years, there is a tendency to use richer text representations than just keywords-based and concepts-based ones in the field of text information processing.

Event originated from cognitive science often appears in the literature of philosophy, cognitive science, linguistics, and artificial intelligence. It has been widely used in the computational linguistics as well as information retrieval and various NLP applications, which plays a special and important role in understanding text semantic. It not only contains specific correlations among a group of text elements but also indicates logical dependencies of things and attracts more and more attentions of researchers. Cognitive scientists believe that event is not only the basic unit of human cognizing and understanding objective world but also storage cell of proposition memory [1]. Most of the current natural language processing technologies lay particular stress on the theory of grammar structure, while ignoring the importance of semantic understanding, especially event semantic understanding [2]. Event-based text representation conforms to the rules of human cognition and natural language understanding.

Seen from present literature on event-based text representation we have consulted, there are the following main problems:

(1) The research on event-based text representation is still in its infancy; the thinking of event network is just beginning to blossom that it is necessary to be further explored.

(2) The operations and applications on event network need to be raised and further researched.

Against the shortcomings of current traditional text representation, the paper takes event as feature item of text and proposes an event-based text representation method. Event is regarded as semantic unit of text, and the events are connected by certain types of relations in the text, and these events imply correlations of linguistic units in the text by making the linguistic units (word, concept, sentence, etc.) as certain elements of event. It no longer regards text as an aggregation of independent words; consequently, the problem of "a bag of words" in classic text representation is

solved. Event network not only keeps semantic information of text and presents events and relations between events but also reflects importance and dynamic behavior of events. Compared to a traditional text representation, event network can express the higher granularity of semantic meaning, closer to the reality and easier for computers simulating text understanding and memorizing of human. It will provide new technology and method for semantic-based text information processing.

The paper is organized as follows: Section 2 introduces the related work. Section 3 constructs event network of Chinese texts in the field of emergencies. Section 4 evaluates the representing effect of event network. In Section 5, the formal definition of event network model for Chinese text is generalized by inducting and abstracting the instances of event network, and then the advantages of the model are analyzed. Finally, we summarize the paper and give an outlook of the future study.

2. Related Work

2.1. The Shortcoming of Traditional Text Representation. In the fields of information retrieval and natural language processing, the traditional text representation models mainly include the following: Boolean model [3], VSM (vector space model) [4, 5], BOW (bag of words) [6], LSI (latent semantic index) [7], LDA (latent Dirichlet allocation) [8], probability retrieval model [8], N-gram model [9], and language model [10].

Semantic information of text is composed of two parts: text component term (word, concept, sentence, etc.) and relationships between terms. Traditional text representation ignored the value of the order and relationships of the component terms on semantic expressing and assumed that the terms are independent, while, in fact, the semantic meaning of text is related to not only component terms and their frequency but also assembly rules and the order of terms, which means that the word-to-word and sentence-to-sentence relationships have an effect on text semantic. The same terms with same frequencies may express different semantic, such as the two following text snippets "Tom gave Mary a book as birthday gift" and "Mary gave Tom a book as birthday gift"; traditional text representation cannot express the difference between them [11]. Text representation based on word unit or concept unit will miss the information of relationships between terms, which will loss semantic meaning of text and result in failing to reflect higher level of semantic information. From the view of event semantic understanding, the above two text snippets express two different events.

In various texts, such as novel, opera, biography, and news reports, that contain many events, traditional text representation did not pay enough attention to event or represent event and relations appropriately. From the perspective of semantic understanding, linguists think that text is not only a group of attributes and concepts but also a describer of a series of events in a higher granularity; according to the thinking, these texts should be regarded as a group of events related by some relations, which is much closer to the laws of human cognition and understanding. From the perspective of formation of text, elementary language units (word, concept, sentence, etc.) form sentence by certain linguistic rules and sentences form a sequence of sentences or paragraph and then form text and express some semantic meaning and theme. Taking event as semantic unit of text and text component term as event-element only solves the problem of "a bag of words" but also expresses the higher level of semantic information.

2.2. Event-Oriented Text Representation. (Although the definitions of event are not unified in different applications, most of them emphasize two kinds of event attributes, action (verb or gerund) and characteristics of action (participant, location, time, etc.), so most researches are centered on verb and attributes of verb. In the paper, attribute of event is called event-element or element for short.) Looking from the current literature we have consulted, little research has been done on event-oriented text representation; the related work mainly includes the following.

Feng [12] proposed incident threading to represent English news reports at sentence level. The texts that describe the occurrence of a real-world happening are merged into a news incident, and incidents are organized in an incident threading by dependencies of predefined types. However, it does not do well in representing Chinese texts.

Glavaš and Šnajder [13] proposed an event-based text representation; however it only has temporal relation. Zhao-Man and Zong-Tian [14] proposed event lattice to represent narrative texts based on concept lattice. In the lattice, text is the object, event is the attribute, and binary relation is used to judge whether an event belongs to a text. Although lattice has precise mathematical properties, its describing power is weak, lacking the ability to express luxuriant relations. And obviously the event lattice has no meaning to one text. It is more suitable to represent inclusion between a group of texts and events than relations between events in a text.

Jian-fang and Yun-yu [15] expounded the thinking of event-based text representation in the paper named "The Research on Event-Oriented Text Representation." This paper discussed the feasibility and adaptability of event-based text representation for Chinese news reports at genre and arrangement of text. However it oversimplified relations between events, resulting in the fact that its representing power is weak. Thus there are still many issues need to be further studied.

Extracting events is the most important thing of event-based text representation. The three main approaches of extracting events are data-driven [16], knowledge-driven [17], and hybrid [18]. The accuracy is about 70 percent according to ACE (Automatic Content Extraction). The paper uses prior knowledge-guided approach.

3. Constructing Event Network of Chinese Text in the Field of Emergencies

Our experimental corpora, CEC (Chinese Event Corpus), are collected from Internet, the texts of which can be divided into five categories: earthquake, fire, traffic accidents, terrorist

Table 1: Statistics of annotated texts.

Text	300
Event	3977
Relation	2023
Coverage of text	85%

attack, and food poisoning according to the classification system of news report about emergency event [19]. Up to now, there are 500 texts in CEC, and 300 ones of them are human annotated event and relations. Some rules have been discovered based on the annotations using mining technology. KBR-EM (knowledge base of rule of event manifestation) has been constructed on CEC.

Verb plays an important role in semantic understanding; it is also core of event. As long as there is verb, it will involve maker and/or receiver of action, and certain regular collocation relationships will be established between action and involved entities; based on this, language would form various basic syntactic configurations and then explain construction of statement and relationship of vocabulary, and so forth. By annotating event on CEC, we find that event corresponds to verb or gerund, and 83% of these verbs or gerunds involve one or two entities, and arrangement of different text typology could affect the layout of events. The relations between events in text are as follows: some are contained in verb of sentence, some are expressed by conjunction (many conjunctions of text virtually show nontaxonomic relation between events, such as "because, therefore" indicates causation), and some are implied in the order of events (such as following relation); the experiment shows that two events will appear successively in text with great probability if there is a relation between them in reality. Our experiments show that events and relations meeting the above findings can cover 85% of entire text. Furthermore, following and causation are the largest number of relations, accounting for 81% of total relations. Statistics of the annotation are displayed in Table 1, where coverage of text is the ratio of event-contained sentences to total sentences. Thus it can be seen that event-based text representation will express text information appropriately.

The guidance of the KBR-EM modified the existing NLP tools (such as tokenizer, part-of-speech tagger, syntactic analyzer, and HowNet), and all of the programs are implemented in Java. Text is processed with word segmentation and POS tagging, syntax analysis and grammatical component tagging, identifying sentence and sentence components, and corresponding sentence or sentence components to event or event elements, regarding verb and gerund as trigger of event. and removing stop-using verbs, such as high-frequency verbs (be, do, have, etc.) and subjective verbs (feel, believe, etc.). Such events belong to stop-using events that are triggered by stop-using verbs; furthermore, stop-using events also include future events and negative events that are triggered by future-tense and negative-form verbs, respectively. Stop-using events should not be included in event network of text. Trigger-associated major components of action are other elements (time, place, subject/predicate-participant, etc.) of the event.

For the identified events, use electronic dictionary and ontology and make concept-climbing after mapping trigger of event into concept. Cluster event and generate event hierarchy by clustering based on the above climbed result, and taxonomic relations between events will be identified. According to the conjunction and other syntactic components of sentences where events are extracted from, consult the findings on relations mentioned above and identify nontaxonomic relations between events.

After identifying events and relations, event network is constructed as follows. Events in the text are arranged in a special directed graph. A named edge from event A to event B means that there is a relation between them in the text, either taxonomic (A is a B, forming multi-inheritance-allowed inheritance diagram) or nontaxonomic, such as causation (A leads to the happening of B), following (A precedes B in time), and composition (A is a part of B). And if there is more than one relation between event A and event B, then one relation is linked to one edge.

4. Experiment and Evaluating Representing Effect

Representing effect could measure whether a text representation method can represent information of original text appropriately and properly. The paper evaluates representing effect of event network with event recall rate (ER), event precision rate (EP), relation recall rate (RR), and relation precision rate (RP).

To compare between events and relations, the paper specifies some rules as follows:

(1) Two events are identical if and only if corresponding event elements are identical that are contained in the individual event.

(2) Two relations are identical if and only if corresponding items are identical that are contained in the individual relation tuple. For taxonomic relation $Is_{-a}(e_u, e_l)$, where e_u is superevent or upper-event, e_l is subevent or lower-event. For directed nontaxonomic relation $r\langle e_1, e_2\rangle$ and undirected nontaxonomic relation $r(e_1, e_2)$, where r is name of the relation, e_1 and e_2 are two events that are connected by the relation

Evaluating on event set of event networks of texts in the field of emergency, as shown in Figure 1, the average recall rate and precision rate are 82% and 88%, respectively. Evaluation of relation set is shown in Figure 2; the average recall rate and precision rate are 76% and 85%, respectively. Compared with previous method [15], the method constructs event network from tagged corpus with events, causation, and following relations, the resulting event network added another adjacent relation and event-element-shared relation. Its event recall and precision rate will be higher, and events contained in the event network can be viewed as complete and correct in theory. According to the findings described in Section 3, nontaxonomic relation recall rate should be at least 81%. However, there are large amount of redundancy

Table 2: Comparison for event network and incident threading.

	ER	EP	Event F_measure	RR	RP	Relation F_measure
Event network	82%	88%	84%	76%	85%	80%
Incident threading	41%	98%	58%	18%	92%	30%

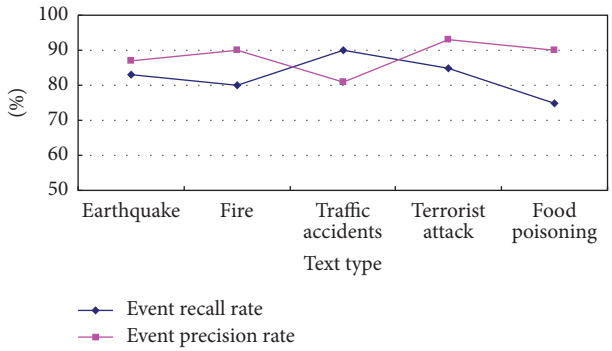

Figure 1: Evaluation of event set.

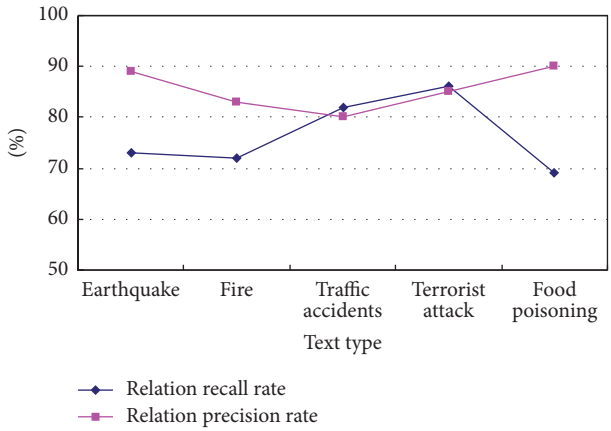

Figure 2: Evaluation of relation set.

and error in adjacent and event-element-shared relation; for example, adjacent relation could actually be following or with no meaningful relation, and event-element-shared relation is too general to specify a relation. So relation precision rate of this method is far inferior to the paper.

Incident threading [12] did well in representing preprocessed grouped English news texts; however, it is less suitable for Chinese text than event network. Evaluating the two representation methods is shown in Table 2.

5. Event Network Model for Chinese Text

An event network contains one or more events that are connected by relations. Events in the network are arranged in a graph, and two events are directly connected by one or more directed/undirected edges (the number of edges depends on the number of relations between the two events) and have some relations. The text representation method is called event network. Though constructing event networks of a large number of texts, we discover that event network is different from general directed digraph. There is information on its each node and each edge, and multiple edges may exist between two nodes. The formal event network model is defined as following by generalizing and abstracting instances of event networks.

Definition 1 (event network). The tuple EN $= (E, R \diamond (R_T, R_{\mathrm{NT}}))$ is called event network that meets the following conditions:

(1) $E = \{e\}$ is nonempty node set, called event set.

(2) $R \diamond (R_T, R_{\mathrm{NT}})$ is edge set, called relation set.

R includes taxonomic relation R_T and nontaxonomic relation R_{NT}. Taxonomic relation $R_T = \{Is_{-a}(e_u, e_l) \mid e_u \in E, e_l \in E\}$ forms multi-inheritance-allowed inheritance diagram, where e_u is superevent and e_l is subevent. R_{NT} forms special graph structure, including directed $R_{\mathrm{NT}} = \{r\langle e_1, e_2\rangle \mid e_u \in E, e_l \in E\}$ and undirected $R_{\mathrm{NT}} = \{r(e_1, e_2) \mid e_u \in E, e_l \in E\}$, where the relation between event e_1 and e_2 is named as r.

Event network can be seen as directed graph. It not only keeps semantic information of text and represents events and relations between events but also reflects importance, dynamic behavior, and state changing of events. Compared with traditional text representation such as VSM, the salient advantage of event network is that it implies correlations among linguistic units of the text in its events, which not only solves the problem of "a bag of words" but also inflects the higher granularity of semantic meaning. Meanwhile relations link events together that can express logical dependencies of things and reflect the occurrence and development process of event.

Event network is a directed graph with information on its nodes and edges. Using all information, various calculations can be done on it by considering some properties of directed graph; for example, an event network can be clustered according to the similarity of events, partitioned into hierarchical structure with different threshold value, and reduced according to importance of event or can keep some other properties. The similarity of texts can be calculated according to the matching of their individual event network; some knowledge can be obtained through mining frequent and simultaneous event elements in multiple event networks. These calculations must meet not only properties of graph but also meaning of information on nodes and edges of event network, so the unique properties and special computation model of event network need to be researched. Establishing abstract operations on event network, some problems will be solved by mathematical methods, which are a kind of good

form for semantic calculation and will support event-based text information processing.

6. Conclusions and the Future Work

The paper introduced requirement of event-based text representation. The formal event network model for Chinese text is defined by abstracting instances of event network on CEC texts. The difference between event network and traditional text representation is that event network keeps semantic information of text, no longer regards text as an aggregation of independent words, and solves the problem of "a bag of words." In addition, it reflects relations among events, importance, and dynamic behavior of event. Our experiments demonstrate the feasibility, adaptability, and advantage of event network as a text representation method.

In the future work, we will study computations on event network by using graph theory, clustering, formal concept analysis, granular computing, and so forth, considering particularity of the model. In this way, various applications of text will be solved by mathematical methods. Theoretical model and method support for present text information processing based on semantic meaning will be provided.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The work is supported by Science and Technology Innovation Project of Chinese Ministry of Culture (2015KJCXXM19) and Foundation for University Youth Teachers of Shanghai (ZZGCD15002).

References

[1] P. Yun-he and W. Geng, "An introduction to intelligent-computing oriented memory theory," *Journal of Computer Research and Development*, vol. 31, pp. 37–42, 1999.

[2] L. Zhong, *The theory of BSCM presented and the implementing of the natural language understanding studied*, East China Normal University, 2004.

[3] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley-Longman, 1st edition, 1999.

[4] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[5] D. D. Lewis, "Evaluation of phrasal and clustered representations on a text categorization task [A]," in *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 37–50, 1992.

[6] Bag-of-words model [DB/OL], http://en.wikipedia.org/wiki/Bag_of_words_model.

[7] T. K. Landauer and M. L. Littman, "Fully automatic crosslanguage document retrieval using latent semantic indexing," in *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pp. 31–38, Waterloo Ontario, 1990, http://www.es.duke.edu/~mlittman/docs/x-lang.ps.

[8] M. David and Y. Blei Andrew, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] T. Chew Lim, S. Sung Yuan, Y. Zhaohui et al., *Text Retrieval from Document Images Based on N-Gram Algorithm*, http://citeseer.nj.nec.com/400555.html.

[10] J. Ponte and W. Croft, "Language Modeling Approach to Information Retrieval," in *Proceedings of SIGIR1998'*, pp. 275–281.

[11] J.-F. Shan, Z.-T. Liu, J.-F. Fu, and Z.-M. Zhong, "Important event extraction of Chinese document based on small world model," in *Proceedings of 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems, ICIS 2009*, pp. 146–150, chn, November 2009.

[12] A. Feng, *Events Threading*, University of Massachusetts, 2008.

[13] G. Glavaš and J. Šnajder, "Event graphs for information retrieval and multi-document summarization," *Expert Systems with Applications*, vol. 41, no. 15, pp. 6904–6916, 2014.

[14] Z. Zhao-Man and L. Zong-Tian, "Events-based Text Similarity Computing," *Journal of Guangxi Normal University (Natural Science Edition)*, vol. 27, no. 1, pp. 149–152, 2009.

[15] Sh. Jian-fang and Sh. Yun-yu, "Research on event network for Chinese text," in *Proceeding of the Interna-tional Symposium on Information Technology Convergence, ISITC2016*, pp. 473–482, Shanghai, China, 2016.

[16] M. Okamoto and M. Kikuchi, "Discovering volatile events in your neighborhood: Local-area topic extraction from blog entries," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5839, pp. 181–192, 2009.

[17] E. Minkov, "Event extraction using structured learning and rich domain knowledge: Application across domains and data sources," *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 2, article no. 16, 2015.

[18] S. Kuptabut and P. Netisopakul, "Event extraction using ontology directed semantic grammar," *Journal of Information Science and Engineering*, vol. 32, no. 1, pp. 79–96, 2016.

[19] Y. Li-ying, L. Hong-juan, and Z. Yong-kui, "The research on classification system of emergency news corpus," in *Proceeding of the monograph of the 25th academic annual conference of Chinese Informatics Association, frontier of Chinese information processing*, Tsinghua University Press, Beijing, 2006.