Journal of Computational Information Systems 11: 13 (2015) 4739–4748
Available at http://www.Jofcis.com

Annotation Event Relation for Chinese Newswire Text Document ⋆

Xianchuan WANG [1,2,*], Zongtian LIU [1], Peitao WEI [1]

[1] *School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China*
[2] *School of Computer and Information Engineering, Fuyang Normal College, Fuyang 236041, China*

Abstract

Event is the coarse-grained form of knowledge representation compared with that of word or concept. Event has more semantic information than word or concept, which is the basic unit of knowing and understanding the real-world for human. There are inherent relations between different events in objective world. At present, there is no relatively full-fledged corpus about event relation in Natural Language Processing field. Therefore, in this paper, we regarded event as the unit of knowledge representation, then we semi-automatically annotated event relation in the aspects of semantic, timing, co-reference and co-participant relation for Chinese newswire news text documents. The purpose of this paper is to supply an evaluation benchmark for event relation detection and platform for text mining. The mean Kappa-score, which illustrated the consistency of annotation, is 93.5%.

Keywords: Event; Event Relation; Newswire Text Document; Annotation

1 Introduction

With the development of the Internet, the number of information explosively become large, especially newswire text document. Information extraction contributes to finding the needed information from the massive text resources. The result of information extraction to some extent depends on structural texts with semantic information, which has requirement that we automatically or semi-automatically annotate semantic information for raw texts. Now, annotation semantic information for texts is still one of research focuses in (Natural Language Processing) NLP field.

Some cognitive scientists believe that event is the basic unit of knowing and understanding, which is the way to describe the objective world for human [1]. When people describe or spread some information, they also regard the event as the basic unit of knowledge representation. For example, the APEC was hold in November 2014 in Beijing. We often describe something like this, when we make a summary or a plan. This methodology, which regards an event (held) as a

⋆Project supported by the National Nature Science Foundation of China (No. 61273328 and No. 61305053).
*Corresponding author.
Email address: xch_wang@shu.edu.cn (Xianchuan WANG).

1553–9105 / Copyright © 2015 Binary Information Press
DOI: 10.12733/jcis14564
July 1, 2015

unit of knowledge representation, with a certain period of time (November 2014), a specific place (Beijing) and other information factors, can record the main factors of an event: what happened, who or what participated, where did it happen and when. Similarly, news agencies often describe the dynamic development process of things based on event in the news texts. This methodology can solve semantic deficiency and the Tennis Problem [2], which were caused by the methodology that word or concept [3, 4] is the basic unit of knowledge representation.

Chinese newswire news text is an expression method that new agency reports event and event relation in real-world with Chinese characters. This also illustrates the fact that there are inherent or objective relations among events in real-world around us. For example, There was an 8.0 magnitude earthquake in Wenchuan county Sichuan province in 2008, which caused thousands of death. The event earthquake and the event died have objective intense causality relation. Therefore, these texts can be regarded as the composition of a series of events based on the inherent or objective relations.

In this paper, we treated event as the basis unit of knowledge representation for text. According to the inherent event relations, we semi-automatically annotated event relations for Chinese newswire news texts in the aspects of semantic, timing, co-reference and co-participant relations. The Chinese newswire news texts with event relations annotated, to some extent, can objectively indicate the process of event happening and developing, which contributes to understanding the contents news reported. These annotated texts can provide a benchmark evaluation on event relation detection and be good to text mining.

In the rest part of this paper, we first gave some definitions about event and event relations. In third section, we indicated the process and methodology of annotating event relations. Next, we did some experiments on CEC and ACE05, and did some analysis on event relation in texts annotated. Finally, we gave a brief survey of related work and drew a conclusion with directions of future work.

2 Definition of Event and Event Relation

Definition 1 *Event: We define event as a thing happening in a certain time and environment, which has some actors and expresses some action features. Event e can be defined as a 6-tuple formally: e ::= def < A, O, T, V, P, L> Elements in 6-tuple is event factors called. A, O, T, V, P, L respectively means Action, Object, Time, Environment, Assertions and Language Expression.*

There are many events (also called natural events) describing what happens in the objective world. However, there are also many events describing the subjective opinion of some persons and institutions or expressing the neutral viewpoint of them, e.g., claim, report and condemn etc. They differ greatly from natural events. Therefore, we introduce the concept of thought event.

Definition 2 *Thought Event: Thought event is the event, in which a certain person has a section thought contents in his or her mind. The thought contents are displayed by oral expressing or characters descripting and keeping in one's mind.*

For example, Xinhua news agency reported the news that there was a magnitude 8 earthquake in Wenchuan county Sichuan province on 12th May, 2012. The event 'report' is a thought event,

and the content of news reported is the thought contents, which described an event happening in real word around our life. The relation between thought event and thought content is defined as thought content relation.

Definition 3 *Event Relation: Event relation means inherent semantic, timing, co-reference and co-participant relations between two events in objective world. It is classified into classification relation and non-classification relation. The latter includes causality relation, composition relation, composition relation, following relation, co-reference relation, co-participant relation.*

Definition 4 *Thought Content Relation: Thought content relation is the event relation that with regard to event A and event B, if event A is a thought event, and event B is the event included the content of event A, then, we call the relation between event A and event B is thought content relation.*

For example, the relation between thought event 'reported' and the event 'earthquake' is thought content relation in example sentence of definition [2].

Definition 5 *Composition Relation: Composition relation is the event relation that with regard to enumerable events, if event A, event B, event C and event D are natural event, and the accomplishment of event A is on the basis of the other three events in certain sequence, then, we call event A is composited of the other three events, the relation between event A and others is composition relation.*

For instance, building house is the composition of laying foundations, building wall and sealing top etc. The start of the latter event is based on the end of the former event among the other three events.

Definition 6 *Causality Relation: Causality relation is the event relation that with regard to event A and event B, if they are natural events, and the occurrence of event A caused the occurrence of event B, then, we call the relation between them is causality relation.*

Definition 7 *Accompany Relation: Accompany relation is the event relation that with regard to two events, in the aspect of time, if the start or the end of event A and event B is synchronous, then, we call the relation between them is accompany relation.*

Definition 8 *Follow Relation: Follow relation is the event relation that with regard to two events, in the aspect of time, if the start of event B to some extent follows the end of event A, then, we call the relation is follow relation between event B and event A.*

Definition 9 *Co-reference Relation: Co-reference relation is the event relation, which is based on the aspect of language expression of event. It means the phenomenon that there are many times description about the same event with the same words or different words.*

For example, on 12th May, 2008, there was an earthquake in Wenchuan county Sichuan province. Vibration sense was violent in Chengdu. The earthquake caused thousands of causalities. There are three events, that is, earthquake, vibration sense and earthquake in the example above sentences. The relation is co-reference relation among the three events.

There are intense semantic similarity between two events, which are of co-reference relation. In this paper, we obtained the co-reference relation between two events by calculating semantic similarity between them. The following Eq. (1) can get the semantic similarity between two events [5], which is based on the HowNet.

$$Sim(W1, W2) = \max(Sim(C_{1i}, C_{2j})), i = 1 \cdots n, j = 1 \cdots m \quad (1)$$

C_{1i} is the n-item word meaning of $W1$. C_{2j} is the m-item word meaning of $W2$.

Definition 10 *Co-participant Relation: Co-participant relation is the event relation based on the objects of event. It means that if event A and event B have the same participant, then, we call the relation is co-participant relation between them.*

3 Annotation Event Relation

In this paper, the Chinese newswire news text documents are regarded as the raw corpus for annotating event relation in emergency domain. Firstly, we automatically crawled these documents from the Internet. Then, pre-processed them, and automatically detected event trigger. Finally, semi-automatically annotated event relation based on domain knowledge engineers. The event relation annotated included causality relation, thought content relation, composition relation, accompany relation, follow relation, co-reference relation and co-participant relation. Fig. 1 indicates the flow chart of annotating event relation.

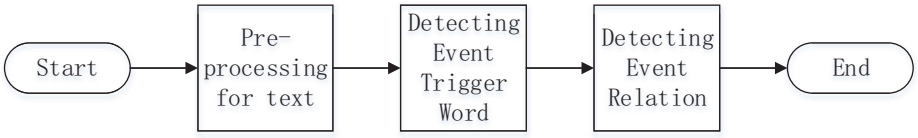

Fig. 1: The flow char of annotating event relation

3.1 Pre-processing text

We crawled Chinese newswire news texts from the Internet with the Ntuch by the key words that are earthquake, food poisoning, terrorist attack, traffic accident and fire. And then, we preprocessed these texts with the linguistics cloud [6] supplied by the agency of HIT and iFLYTEK. The process of preprocessing contained segmenting word and tagging lexical category.

3.2 Event detection

Event detecting is a key step for annotating event relation. In this paper, detecting event means detecting event trigger word. Event trigger word is the word in sentence, which can mostly illustrate the occurrence of an event. Event trigger word and the occurrence of an event have relation, which is one to one relation. We detected trigger word with the trigger word-table, which is constructed with the annotated Chinese newswire news texts. The trigger word-table is

called $T1$. Then we expanded $T1$ with Chinese synonyms dictionary and obtained trigger word-table expanded we called $T2$. Next, we extracted verbs from the texts preprocessed, and saved these verbs to list $L1$. Finally, we detected these verbs extracted by Algorithm 1, which is called matching algorithm by trigger word-table for detecting event. After matching event trigger word, we manually detected the undetected event trigger word based on domain knowledge engineers, which could elevate the precision of detecting event trigger word. The step as follows:

Step 1 For each $verb_i$ in $L1$ Do;

Step 2 Search with $verb_i$ in $T2$;

Step 3 If $verb_i$ in $T2$;

Step 4 $L2$=$L2$ U $verb_i$.

3.3 Annotation event relation

In this paper, on the basis of detecting event trigger word, we semi-automatically annotated event relation, which is based on domain knowledge engineers. There are two important steps to annotate event relation. Firstly, these knowledge engineers judged whether there is a relevance or correlation between two events. Then, they confirmed the specific type of event relation between the two events. According to the priority, which was semantic relation, timing relation, co-reference relation and co-participant relation, we detected event relation and annotated in the aspect of chapters. The semantic relation included causality relation, composition relation and thought content relation. The timing relation contained accompany relation and follow relation. The annotation step as follows:

Step 1 For two trigger words, t_i and t_j in $document_i$ do;

Step 2 If Correlation(t_i,t_j) do;

Step 3 If Relation(t_i,t_j)do;

Step 4 Relation=Relation U Relation(t_i,t_j).

With regard to the texts detected trigger word, the following is the concrete process of annotating event relation. Firstly, different domain knowledge engineers got the same Chinese newswire news texts. Then, these knowledge engineers independently annotated the same text assigned. Thirdly, the conflicting event relation annotated is submitted to symposium, in which there were many persons taking part. And, they confirmed the final event relation in order to keep annotation consistency. Finally, we verified the correctness of these texts annotated with XML format by some codes.

4 Experiment and Evaluation

4.1 Datasets and evaluation method

We respectively selected CEC and ACE05 Chinese edition as the datasets for our experiments on trigger word detection. CEC is a Chinese event corpus in emergency domain, which includes five

topics, they are terrorist attack, traffic accident, food poisoning, fire and earthquake. The CEC has more than 330 articles annotated. We respective randomly selected 20 articles and 40 articles from each topic in CEC and ACE 05.

Table 1: Datasets for trigger word detection

Item	Training Datasets	Testing Datasets
CEC	100	200
ACE 05	320	300
Gross	420	500

As can be seen from Table 1, the gross number of training datasets respective is 100 articles and 320 articles and the gross amount of testing datasets respective is 200 articles and 300 articles from CEC and ACE 05. The approach we use for measuring the agreement of detecting event trigger word is the universal method, which is the precision, recall and F-measure.

4.2 Experiment result and analysis

Table 2 shows some data of precision, recall and F-measure for detecting event trigger word in CEC and ACE 05 by the methodology we used, which respective is trigger word-table and trigger word-table expanded.

Table 2: Event trigger word detected

Corpus	Trigger Word-Table			Trigger Word-Table Expanded		
	P	R	F	P	R	F
CEC	62.3	65.1	63.7	75.8	72.7	74.2
ACE05	59.4	61.2	60.3	72.3	69.5	70.9
Mean	60.85	63.15	62	74.05	71.1	72.55

As we can see from the above Table 2, the mean F-measure of detecting trigger word is 62% with trigger-table in CEC and ACE 05. However, the mean F-measure of detecting trigger word is 72.55% with expanded trigger-table in CEC and ACE 05. The latter is elevated by 10.55% compared with the former. Actually, the trigger word-table, which is constructed from training datasets, is regarded as domain background knowledge for detecting event trigger word in testing datasets. Chinese is a paratactic language, which cause the phenomenon that there are many different event trigger words describing the same event. This is the reason why the expanded trigger word-table is priority to the trigger word-table in the aspect of mean F-measure. In the respect of corpus, the detecting effect in CEC is elevated by more than 3.5% compared with that in ACE 05. The reason is that ACE 05 is annotated based on predicate. However, there are many predicates, which expresses the state of some semantic role of event. These predicates are not events to some extent. Trigger word-table depends on the annotated texts with constructed format, which indicated the facts that the methodology we used is based on some domain knowledge in this paper.

Causality relation, composition relation and thought content relation primarily indicate the semantic relation among events. Accompany relation and follow relation mainly illustrate timing relation among events. Co-reference relation describes event relation in the aspect of language expression of event. Co-participant expresses event relation in the respect of object factor of event. Table 3 indicated the distribution of event relation we annotated for Chinese newswire news texts.

Table 3: Distribution of event relation annotated

Type		EQ	FR	FP	TAT	TAK	Gross
Semantic Relation	Causality	247	87	97	351	135	917
	Composition	1	7	0	1	1	10
	Thought Content	168	109	92	155	165	689
Timing Relation	Accompanying	30	72	77	276	95	550
	Following	40	131	117	323	147	758
Co-reference		258	285	226	263	181	1213
Co-participant		112	87	64	125	57	447
Gross		856	778	673	1495	783	4585

The abbreviated words that they are EQ, FR, FP, TAT and TAK in line 1 of Table 3 respectively stand for earthquake, fire, food poisoning, traffic accident and terrorist attack. The Table 3 gives us some data about distribution of event relation we annotated. As we can see from the above table, there are 4 super categories event relation, that is, semantic relation, timing relation, co-reference relation and co-participant relation. Additionally, the semantic relation includes causality relation, composition relation and thought content relation. The timing relation contains accompany relation and follow relation. We assigned the priority for the four super categories event relation as they displayed order in the upper table, and according to this priority, we annotated event relation for Chinese newswire news texts.

From the upper table, we can find the distribution of event relation we annotated. From more to less, the number-order of event relation is co-reference relation, causality relation, follow relation, thought content relation, accompany relation and co-participant relation and the last one is composition relation. There are 1213 co-reference relations, which are the largest among these event relations we annotated. And the number of composition relation we annotated is only 10, which is the least.

The general amount of these event relations is 4585. Under the circumstance, without taking the transitivity between event relations into account, the event relation we annotated almost can cover all the event detected in Chinese newswire news texts. According to the number of co-reference relation and co-participant relation, we verified the phenomenon about default semantic information described, which accord with human communication habits. The approach we use for measuring the consistency of event relation we annotated is Kappa-score, which is calculated by the following formula

$$K = \frac{P_o - P_c}{1 - P_c}. \tag{2}$$

P_o stands for consistency annotated: the percentage of result consistency annotated.

P_c stands for desiring consistency: the percentage of expected result consistency annotated.

Table 4: Event relation annotated kappa-score

Item	Event Relation K(%)
Earthquake	91.6
Fire	92.8
Food Poisoning	95.3
Traffic Accident	93.2
Terrorist Attack	94.5
Mean	93.5

Table 4 indicates some data about the Kappa-score of event relation we annotated for Chinese newswire news texts. As we can see from the upper table, we can find the facts that the Kappa-score is 91.6% in the topic of earthquake, which is the least. And the Kappa-score in the topic of food is the largest, which is 95.3%. The mean Kappa-score of the five topics is 93.5%. Although, there is some difference in respect of Kappa-score, the largest gap is 3.7% in Kappa-score among these five topics. The event relation domain knowledge engineers semi-automatically annotated has good consistency annotated.

We semi-automatically annotated event relation for Chinese newswire news texts, which is with the help of annotation tools we developed. And the methodology we use for annotating is based on domain knowledge engineers. Therefore, the annotation precision and consistency is affected by the background knowledge and experience of domain knowledge engineers. There is some limit in this paper, but our work is very meaningful for information extraction, especially, event relation extraction. Our work is the fundamental work for the research interest of event relation.

5 Related Work

There have been many research work in the field of event detection, event extraction and application [7-9] since MUC-6 [10] initiated the tasks about information extraction in the aspect of event. However, at present, the research interest in event relation is in a fledging period all over the globe. There is no full-fledged linguistics resource about event relation [11].

Linguistic Data Consortium (LDC) released TimeBank1.2 [12] in 2003, which annotated event, time and the relation between event and time for 300 English newswire news texts. TimeBank provided corpus support for studying on event timing relation. Next year, LDC started ACE tasks, which treated event extraction as one of the tasks in information extraction. And LDC released ACE 05 evaluation corpus in 2005, which were multi-language training resources including Chinese edition, English edition and Arabic edition. There were many relations in ACE 05, which mainly indicated entity-relation and the relation between event and state. Both of TimeBank and ACE 05 annotated relation with regard to specific tasks, and the two corpuses were based

on predicate. However, there are many predicates in newswire news texts, which described the state of some factors of event. Therefore, these predicates are not events indeed.

The paper [13] annotated causality relation between event instances with dependency parsing, when they did some work on syntactic parsing for Japanese sentences. Fu [14] annotated causality relation between two events for 200 Chinese newswire news texts in the respect of inner-sentence, sentence and paragraph. These texts he annotated included five topics in emergency field, which were fire, earthquake, food poisoning, terrorist attack and traffic accident. The causality relation he annotated contained one cause to one effect, one cause to many effects, many causes to one effect and many causes to many effects. Ma [15] did some research work about the correlation among events, and manually annotated event relation for Chinese newswire news texts collected, in which there were six topics and 180 articles. The event relation annotated was only bi-relation in the respect of correlation, which only included relevant relation and irrelevant relation. The gross number of event-pair he annotated is 2842, in which there were 811 relevant event-pairs. Liao [16] represented text knowledge with event Co-occurrence network structure.

At present, as we expressed above, the number of full-fledged linguistics resources about event relation is very little. With regard to Chinese, there is hardly any linguistics resource to some extent. There are some shadows in existing linguistics resources about event relation, which is that the event relation annotated is relatively simple. For example, Fu [14] only annotated causality relation between two events. And Ma [15] only annotated bi-relation for events. TimeBank [12] only has timing relation. There are complex semantic relation among events in the real world around us, but the event relations annotated in the above linguistics corpuses is far from describing real relation among these events.

6 Conclusion and Future Work

Recently, following fission increasing in the number of texts on the Internet, text mining based on event has been one of research focuses in NLP field. Event is a coarse-grained unit of knowledge representation compared with that of word or concept, which has more semantic information than word or concept. Event is also the basic unit for human to know and understand the objective world. Event is not independent in the real-world around us. There are some objective and inherent connection among events. In this paper, we regarded event as the unit of knowledge representation. And we semi-automatically annotated event relation for more than 300 Chinese newswire news texts in the aspect of semantic, timing, co-reference and co-participant. The event relation we annotated included seven categories, which were causality relation, composition relation, thought content relation, accompany relation, follow relation, co-reference relation and co-participant relation. The gross number of event relation we annotated were more than 4500. These event relations we annotated almost could cover all the event existing in Chinese newswire news texts. And, to some extent, these event relations we annotated could indicate the dynamic process of event happening and developing, which contributed to understanding the contents of texts for human. The mean Kappa-score was 93.5%, which illustrated the result that the event relation we annotated has good consistency and quality.

The event relation we annotated will construct a good platform for text mining based on event. In the future, we will do some applied research work on the basis of the event relation we annotated, for example, event-oriented text representation, automatic summarization, information retrieval and so on.

Acknowledgements

We give sincere thanks to the persons, who jointly with us finished annotating event relation. This research was supported by the National Natural Science Foundation of China under Grant No. 61273328 and No. 61305053.

References

[1] Zongtian L., Meili H., Wen Z., Zhaoman Z., Jianfeng F., Jianfang S., Huilai Z., Research on Event-oriented Ontology Model, Computer Science, 36 (2009) 189-192, 199.

[2] De Mier A., Noy M., A solution to the tennis ball problem, Theoretical Computer Science, 346 (2005) 254-264.

[3] Van den Broek P., Using Texts in Science Education: Cognitive Processes and Knowledge Representation, Science, 328 (2010) 453-456.

[4] Chen L., Fan R.X., Gao Q., Model of text representation based on concept, Computer Engineering and Applications, 44 (2008) 162-164.

[5] Feng L.I., Fang L.I., An New Approach Measuring Semantic Similarity in Hownet 2000, Journal of Chinese Information Processing, 21 (2007) 99-105.

[6] Che W., Li Z., Liu T., Ltp: A chinese language technology platform, In: Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, Association for Computational Linguistics, (2010) 13-16.

[7] Glavas G., Snajder J., Event graphs for information retrieval and multi-document summarization, Expert Systems with Applications, 41 (2014) 6904-6916.

[8] Ananiadou S., Pyysalo S., Tsujii J., Kell D.B., Event extraction for systems biology by text mining the literature, Trends in Biotechnology, 28 (2010) 381-390.

[9] Chang C.H., Kayed M., Girgis M.R., Shaalan K.F., A survey of web information extraction systems, Ieee Transactions on Knowledge and Data Engineering, 18 (2006) 1411-1428.

[10] Grishman R., Sundheim B., Message Understanding Conference-6, In: A Brief History COLING, (1996) 466-471.

[11] Yang X., Ma B., Hong Y., Yao J., Zhu Q., A survey of Linguistics Resource, Evaluation and the Research in Event Relation Detection Intelligent Computer and Applications, (2014) 5-9.

[12] Pustejovsky J., Hanks P., Sauri R., See A., Gaizauskas R., Setzer A., Radev D., Sundheim B., Day D., Ferro L., The timebank corpus, Corpus linguistics, 40 (2003) 647-656.

[13] Abe S., K. Inui, Y. Matsumoto., Two-phased event relation acquisition: coupling the relation-oriented and argument-oriented approaches, In: Proceedings of the 22nd International Conference on Computational Linguistics, 1 (2008) 1-8.

[14] Fu J., Stduy on Knowledge Processing based on Event, Phd Thesis, School of Computer Technology and Engineer, Shanghai University, June 2010.

[15] Ma B., Hong Y., Yang X., Yao J., Zhu Q., Using Event Dependency Cue Inference to Recognize Event Relation, Acta Scientiarum Naturalium Universitatis Pekinensis, 49 (2013) 109-16.

[16] Liao T, Liu Z, Xuan X., Research on Event Co-occurrence Network Structure Based Method for Chinese Text Representation, Journal of Computational Information Systems, 9 (2013): 5535-5542.