Ontology-Based Event Modeling for Semantic Understanding of Chinese News Story

Wei Wang[1,2] and Dongyan Zhao[1]

[1] Institute of Computer Science & Technology, Peking University, Beijing, China
[2] Department of Electronic Technology, Engineering University of CAPF, Xi'an, China
{wjwangwei,zhaodongyan}@pku.edu.cn

Abstract. Describing and extracting event semantic information is essential to build an event knowledge base and applications in event and semantic level. However, most existing work deals with documents so they fail to provide sufficient semantic information about events in news articles. In this paper, considering ***What, Who, When, Where***, **W**hy and ***How***, the 5W1H elements of a piece of news, we propose a News Ontology Event Model (NOEM) which can describe 5W1H semantic elements of an event. The model defines concepts of entities (time, person, location, organization etc.), events and relationships to capture temporal, spatial, information, experiential, structural and causal aspect of events. A comparison to existing event models and an empirical case study show that NOEM can effectively model the semantic elements of news events and their relationship; and has a strong ability to represent knowledge facts and easily adapt to new domains.

Keywords: Ontology, 5W1H, Event Model, Ontology Population.

1 Introduction

In the past decade, event has emerged as a promising research field in Natural Language Processing (NLP), Information Retrieval (IR) and Information Extraction (IE). In online news services domain, event-based techniques which can extract entities, such as time, person, location, organization etc. from news stories and find relationship among them to represent structural events, have been paid wildly attention. Based on the identification and extraction of these valuable event facts, more convenient and intelligent services can be implemented to facilitate online news browsing in event and semantic level.

However, the definition and representation of event are different in various research areas. In the literature, there are a number of event models nowadays. For example, event templates used in Message Understanding Conference (MUC) [1], a structural event representation in Automatic Content Extraction (ACE) [2], a generic event model E [3] [4] in event-centric multimedia data management, and ontology-based event models, such as ABC [5], PROTON [6] and Event-Model-F [7] in knowledge management. But these models are not suitable for semantic

M. Zhou et al. (Eds.): NLPCC 2012, CCIS 333, pp. 58–68, 2012.

understanding of news events. In order to support semantic applications, for example, event information extraction and semantic relation navigation, we propose a News Ontology Event Model (NOEM) to describe entities and relations among them in news events.

As we know, 5W1H (including What, Who, When, Where, Why, How), a concept in news style is regarded as basics in information gathering. The rule of the 5W1H originally states that a news story should be considered as complete if it answers a checklist of 5W1H. The factual answers to these six questions, each of which comprises an interrogative word: what, who, when, where, why and how, are considered to be elaborate enough for people to understand the whole story [8].

In NOEM, in order to address the whole list of 5W1H, we define concepts of entities (time, person, location, organization etc.), events and relationships to capture temporal, spatial, information, experiential, structural and causal aspect of events. The comparison of NOEM with existed event models shows that it has a better knowledge representation ability, feasibility and applicability. By automatically extracting structural 5W1H semantic information of events and populating these information to NOEM, an event knowledge base can be built to support event and semantic level applications in news domain.

The rest of the paper is organized as follows. We first review related work in Sec. 2 by discussing event definitions and event modelings. The proposed ontology event model NOEM is introduced in Sec. 3. In Sec. 4, we evaluate the representative ability of NOEM by comparing it with existing event models and demonstrate its feasibility and applicability by means of case study. In Sec. 5, we conclude this paper.

2 Related Work

2.1 Event Definitions

The notion of an *event* has been widely used in many research fields related to in natural language processing, although with significant variance in what exactly an event is. A general definition of *event* is "something that happens at a given place and time", according to WordNet [9]. Cognitive psychologists look events as "happenings in the outside world", and they believe people observe and understand the world through event because it is a suitable unit in accordance with aspect of human cognition. Linguists have worked on the underlying semantic structure of events, for example, Chung and Timberlake (1985) stated that "an event can be defined in terms of three components: a predicate; an interval of time on which the predicate occurs and a situation or set of conditions under which the predicate occurs." In [10], Timberlake further supplements it as "events occur in places, under certain conditions, and one can identify some participants as agents and some as patients."

In recent years, some empirical research have been developed on the cognitive linguistics theoretical basis. TimeML [11] is a rich specification language for event and temporal expressions in natural language text. Event is described as "a cover term for situations that happen or occur. Events can be punctual or last for a period of time". In event-based summarization, Filatova et.al. [13] define a concept of *atomic event.* Atomic events link major constituent parts (participants, locations, times) of events through verbs or action nouns labeling the event itself. In IE community, an event represents a relationship between participants, times, and places. The MUC extracts prespecified event information and relates the event information to particular organization, person, or artifact entities involved in the event. The ACE describes event as "an event involving zero or more ACE entities, values and time expressions".

The event extraction task in ACE requires that certain specified types of events that are mentioned in the source language data be detected and that selected information about these events be recognized and merged into a unified representation for each detected event. According to the requirements of semantic understanding of news and the characteristics of news story, we define event as "an event is a specific occurrence which involves in some participants". It has three components: a predicate; core participants, i.e., agents and patients; auxiliary participants, i.e., time and place of the event. These participants are usually named entities which correspond to the *what, who, whom, when, where* elements of an event. The relationships among entities and events are also concerned. By analyzing the connections between the predicates, we can get the *why* and *how* elements which are cause and effect of the event.

2.2 Event Modeling

Event modeling involves event definition, event information representing and storing. There have been several event models in different application domains.

Probabilistic Event Model. In [12], a news event probabilistic model is proposed for Retrospective news Event Detection (RED) task in Topic Detection and Tracking (TDT). In the model, news articles are represented by four kinds of information: *who* (persons), *when* (time), *where* (locations) and *what* (keywords). Because news reports are always aroused by news events, a news event is modeled by mixture of three unigram models for persons, locations and keywords and one Gaussian Mixture Model (GMM) model for timestamps.

Atomic Event Model. In event-based summarization, Filatova et.al. [13] denote atomic events as triple patterns $<n_m, t_i, n_n>$. The triples consist of an event term t_i and two named entities n_m, n_n. This event model was adopted by a number of work in event-based summarization [14] [15].

Structural Event Model. In IE domain, event model is a structural template or frameset in MUC and ACE respectively. In MUC, the event extraction task is a slots filling task for given event templates. That is, extracting pre-specified event information and relating the event information to particular organization, person, or artifact entities involved in the event. In ACE, the event is a complex event structure involving zero or more ACE entities, values and time expressions.

Generic Event Model. Jain and Westermann propose a generic event model E for event-centric multimedia data management applications [3]. The model is able to capture temporal aspect, spatial aspect, information aspect, experiential aspect, structural aspect and causal aspect of events [4].

Ontology Event Model. Ontology is an explicit and formal specification of a shared conceptualization [16]. It is an important strategy in describing semantic models. ABC ontology [5] developed in Harmony Project[1] is able to describe event-related concepts such as event, situation, action, agent, and their relationships. PROTON, a base upper-level ontology developed from KIMO Ontology in Knowledge and Information Management (KIM) [6] platform, has the ability to describe events which cover event annotation types in ACE. In [7], a formal model of events Event-Model-F is proposed. The model can represent arbitrary occurrences in the real world and formally describe the different relations and interpretations of events. It actually blends the six aspects defined for the event model E and interrogatives of the Eventory system [17] to provide comprehensive support to represent time and space, objects and persons, as well as mereological, causal, and correlative relationships between events.

We mainly concern about ontology-based event models in this paper.

3 News Ontology Event Model

On the basis of analyzing existing event models, we build NOEM for semantic modeling news event in this work. The main advantages of ontology-based modeling lie in two aspects: 1) It is able to provide common comprehension of domain knowledge, determine commonly recognized terminologies, and implement properties, restrictions and axioms in a formulated way at different levels within certain domains. 2) Implicit knowledge can be acquired from known facts (events, entities and relations) by using inference engine of ontology.

3.1 The Design of NOEM

Our goal of designing NOEM is to provide a basic vocabulary for semantic annotation of event 5W1Hs in news stories. So classes and properties are carefully selected to guarantee NOEM's compactedness as well as to supply abundant semantics. In accordance with Jain's generic event model, our model also tries to capture temporal, spatial, information, experiential, structural and causal aspect of events. The proposed event model is able to cover information of events in three levels.

- ***Event information:*** Based on existing event models, we select general concepts such as 'space', 'time', 'events', 'objects', 'agents', etc. to represent an event and its *5W* elements. The ACE event hierarchy is imported to identify event's types by trigger words. We only capture actions which can

[1] The Harmony Project, `http://metadata.net/harmony`

uniquely identify an event. Since events are spatial and temporal constructs, the event information component necessarily contains the time period of the activity and its spatial characteristics. Additionally, entities like people or objects that participate in an event are described. Concepts defined in NOEM is shown in Table 1.

- ***Event relations:*** Events (and the activities underlying them) may be related to other events (activities) that occur in the system. Examples of such relations can be temporal and spatial co-occurrences, temporal sequencing, cause-effect relations, and aggregations of events. By defining new concepts, for example, 'situation' and 'constellation', and properties such as 'precedes' and 'follows', the model achieves the ability of describing an event in a fine-grained manner, and of relating and grouping events. Relations defined in NOEM is shown in Table 2.
- ***Event media:*** Events can be described in various media, e.g. audio, video, text. Since we only care about news stories, we define concepts of 'document' and 'topic' to capture the characteristics of the news articles. Information such as news types, resource locators, or indexes corresponding to specific document that support the given event are modeled. CNML (Chinese News Markup Language) is imported to represent a news article's topic so that we can connect an event to its category in document-level.

Table 1. Concepts in NOEM

| Thing | Entity | Happening | Time | Place| Document | Topic | Phisical | Abstracts
| Event | Situation | Action | Constellation | Agent | LogicalTime | PhysicalTime |
RelativeTime | Logical Place | Physical Place | Relative Place |

Table 2. Relations in NOEM

| hasSubject | hasObject | hasCause | hasResult | isSubeventOf | involvesIn | atTime|
inPlace | predeces | follows | hasAction | describedIn | hasTopic | hasClass | hasType |

3.2 Main Concepts and Properties in NOEM

In this section, we discuss main concepts, properties of NOEM and how they are used to represent 5W1H semantic elements of an event in detail. The designed News Ontology Event Model is shown in Fig. 1, for the sake of clarity, only main concepts and properties are included.

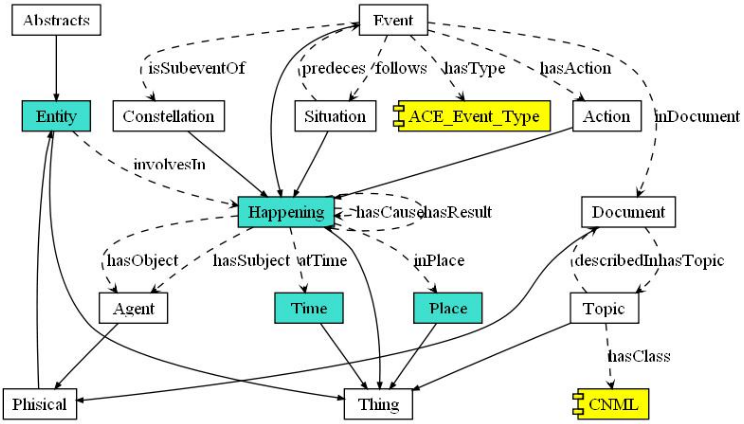


Fig. 1. News Ontology Event Model

- ***Happening*** is the superclass of all types of eventuality. It has four subclasses: 'Event' denotes dynamic event, 'Situation' denotes static status, 'Action' denotes an activity and 'Constellation' denotes an event set. Properties 'hasCause' and 'hasEffect' of 'Happening' represent cause-effect connections between events. This is helpful for future work of analyzing *why* among events.
- ***Event*** is a concept denotes dynamic 'Happening'. An event always has a type which can be used to specify *what*. We import all types of ACE event, 'Business', 'Conflict', 'Contact', 'Justice', 'Life', 'Movement', 'Personnel' and 'Transaction'. They appear as a component in Figure 1.
- ***Situation*** describes static status preceding or following an event. Properties 'precedes' and 'follows' can be used to represent *how*.
- ***Constellation*** is a set of happenings with some relations among them, e.g., cause-effect and core-peripheral. It describes a complete happening caused by a key event and developed by sub-events.
- ***Agent*** is a concept to represent *who* and *whom*. It is the superclass of 'Group' and 'Person'.
- ***Time*** apparently represents *when*. Its subclasses 'logicalTime', 'physicalTime' and 'relativeTime' are reserved for our future work of time normalization and inference.
- ***Place*** represents *where*. Its subclasses 'logicalPlace', 'physicalPlace' and 'relativePlace' are reserved for our future work of location normalization.
- ***Document*** is the media aspect of an event. It has an URI (Universal Resource Identifier) refers to a news article.
- ***Topic*** is a concept in document level. It is related to category of a news story, for example, 'Sports', 'Law', 'Politics' and so on. We import 2082 subclasses from CNML taxonomy for topic classification.

4 Evaluation

Janez Brank et. al. [18] classified ontology evaluation methods into four categories: (1) Comparing the ontology to a "golden standard"; (2) Using an ontology in an application and evaluating the results; (3) Comparing with a source of data about the domain to be covered by the ontology; (4) Evaluation is done by humans who try to assess how well the ontology meets a set of predefined criteria, standards, requirements.

In this section, we evaluate the representative ability, the feasibility and applicability of NOEM using a combination of above methods.

4.1 Comparison with Existing Event Models

When designing the NOEM, we analyzed existing event-based systems and event models with respect to the functional requirements. These models are motivated from different domains such as the Eventory system for journalism, the structural representation of event in MUC and ACE, the event model E for event-based multimedia applications and Ontology-based models, such as the Event Ontology as part of a music ontology framework, ABC, PROTON for knowledge management and Event-Model-F for distributed event-based systems. All models are domain-dependant and they are too simple or too complicated for news event understanding task in this paper.

By analyzing the representative ability of existing Event models, we obtain six factors: action, entity, time, space, relations (here we concern structural, causal and correlation among events) and event media. An overview of the analysis results and comparison to the features of NOEM along the representative ability is listed in Table 3. It shows that NOEM has a better representative ability than structural event models, i.e., probabilistic, atomic and MUC/ACE models. In addition, NOEM's representative ability is as good as PROTON and Event-Model-F, two classic ontologies. At the same time, NOEM has a more compact design with only a few classes and relations, and is suitable to modeling Chinese News.

4.2 Manually Evaluation

To evaluate NOEM in a practical environment, four postgraduate students are invited to manually analyze 6000 online news stories from XinHua and People news agency. These news stories cover 22 topics of CNML such as politics, economy, military, information technology, sports and so on. With the help of headline of each news item, one topic sentence that contains the key event is identified. Then 5W1H elements of the key event are labeled from the headline and the topic sentence according to the NOEM definition. The annotation result shows that 85 percent of online news story can be described by NOEM appropriately.

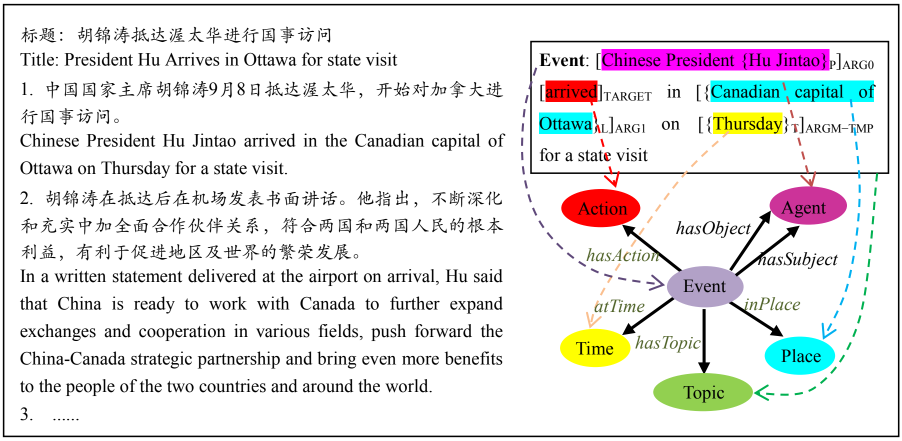
Table 3. Comparison of Event Models

Event Model	Domain	Representative Ability							
						Relation			
		Action	Entity	Time	Space	Str.	Cau.	Cor.	Media
Probabilistic	TDT/RED	Y	Y	Y	Y	N	N	N	N
Atomic	Event-based Sum.	Y	Y	N	N	N	N	N	N
MUC/ACE	IE	Y	Y	Y	Y	Y	N	N	N
Eventory	Journalism	Y	Y	Y	Y	Y	Y	N	N
E	Multimedia Mana.	Y	Y	Y	Y	Y	Y	Y	Y
Event Ontology	Music Mana.	Y	Y	Y	Y	Y	Y	Y	N
ABC	Knowledge Mana.	Y	Y	Y	Y	Y	Y	Y	N
PROTON	IR	Y	Y	Y	Y	Y	Y	Y	Y
F	Event-based Sys.	Y	Y	Y	Y	Y	Y	Y	Y
NOEM	EE	Y	Y	Y	Y	Y	Y	Y	Y

Abbreviations: Str.=Structural, Cau.=Causal, Cor.=Correlation,
Mana.=Management, Sum.=Summarization, Sys.= System

4.3 A Case Study

Here we take a story from Xinhua news agency September 9, 2005 as an example to extract and describe the key event elements. The snippet of the news is shown in the left part of Fig. 2. We first use a machine learning method in our previous work [19] to identify the topic sentence about the key event of this story. And then we use a verb-driven method, along with Name Entity identification and semantic role labeling method proposed in work [20] to get the key event's 5W1H information from the story.



标题：胡锦涛抵达渥太华进行国事访问

Title: President Hu Arrives in Ottawa for state visit

1. 中国国家主席胡锦涛9月8日抵达渥太华，开始对加拿大进行国事访问。

Chinese President Hu Jintao arrived in the Canadian capital of Ottawa on Thursday for a state visit.

2. 胡锦涛在抵达后在机场发表书面讲话。他指出，不断深化和充实中加全面合作伙伴关系，符合两国和两国人民的根本利益，有利于促进地区及世界的繁荣发展。

In a written statement delivered at the airport on arrival, Hu said that China is ready to work with Canada to further expand exchanges and cooperation in various fields, push forward the China-Canada strategic partnership and bring even more benefits to the people of the two countries and around the world.

3. ......

Event: [Chinese President {Hu Jintao}$_{P}$]$_{ARG0}$ [arrived]$_{TARGET}$ in [{Canadian capital of Ottawa}$_{L}$]$_{ARG1}$ on [{Thursday}$_{T}$]$_{ARGM-TMP}$ for a state visit

Fig. 2. Snippet of a news story, the identified key event and its 5W1H elements

From the key event of the news, “Chinese President Hu Jintao arrived in the Canadian capital of Ottawa on Thursday for a state visit.”, we get an ACE “*Movement*” event and its 5W1H elements (right part of Fig. 2). The extracted 5W1H semantic event information, their types and semantic relations are denoted in RDF (Resource Description Framework)[2] triples. For the example event, we get triples <*Chinese President Hu Jintao, arrive, Canadian capital of Ottawa*>, <*arrive, isTypeof, Movement/Transport*>, <*Chinese President Hu Jintao, isTypeof, Person*>, <*2005-09-08T00:00:00, isTypeof, Time*> (*Thursday* is normalized as 2005-09-08) and <*Ottawa, isTypeof, Place*>.

4.4 Population of NOEM

The proposed NOEM is built on Protégé[3]. By using a predefined template, an OWL (Web Ontology Language)[4] file is automatically generated in which triples are mapped to the concepts and relations according to NOEM. By this means, the event 5W1H elements can be populated into Protégé as instances.

A snippet of automatic generated OWL file of the example story is listed below.

```
<Event rdf:ID="NewsEvent_588">
   <rdfs:comment>
      Hu Arrives in Ottawa.
   </rdfs:comment>
   <inDocument>
      <Document rdf:resource="#588"/>
   </inDocument>
   <hasAction>
      <Action rdf:ID="<Chinese President Hu Jintao,arrive,
                        Canadian capital of Ottawa>"/>
    </hasAction>
   <hasSubject>
      <Agent rdf:ID="Chinese President Hu Jintao"/>
   </hasSubject>
   <hasObject>
      <Agent rdf:ID="Canadian capital of Ottawa"/>
   </hasObject>
   <hasType>
      <ACE_Event_Type rdf:resource="#Movement/Transport"/>
   </hasType>
   <atTime>
      <Time rdf:ID="2005-09-08T00:00:00"/>
   </atTime>
   <inPlace>
```

[2] http://www.w3.org/TR/REC-rdf-syntax/

[3] http://protege.stanford.edu/

[4] http://www.w3.org/TR/owl-ref/

```
    <Place rdf:ID="Ottawa"/>
    <Place rdf:ID="Canada"/>
  </inPlace>
</Event>
```

Besides the key event, <*Chinese President Hu Jintao, deliver, a written statement* > is also identified as a subevent and associated to the key event by relationship “follows”. The automatically mapping and populating the 5W elements and relations into Ontology shows the feasibility and applicability of NOEM.

5 Conclusions and Future work

In this paper, NOEM, an event Ontology which describes concepts of 5W1H event semantic elements and relationships of events is proposed. NOEM is able to capture temporal, spatial, information, experiential, structural and causal aspect of an item of news. By taking advantage of logical reasoning ability of the NOEM ontology, the output of *why* and *how* elements together with relationships among *who*, *what*, *whom*, *when* and *where* of events can be used to build a multidimensional news event network. This will largely facilitate online news browsing in event and semantic level.

Our future work is to build a news events knowledge base and a semantic retrieval engine on NOEM. This will strongly support semantic information retrieval on event level and other event level semantic applications.

Acknowledgment. This work is supported by the National High-Tech Project of China (Grant No. 2012AA011101).

References

1. Chinchor, N., Marsh, E.: MUC-7 Information Extraction Task Definition (version 5. 1). In: MUC-7 (1998)
2. ACE (Automatic Content Extraction).: Chinese Annotation Guidelines for Events. National Institute of Standards and Technology (2005)
3. Westermann, U., Jain, R.: E - A generic event model for event-centric multimedia data management in eChronicle applications. In: The 2006 IEEE International Workshop on Electronic Chronicles, Atlanta, GA (2006)
4. Westermann, U., Jain, R.: Towards a Common Event Model for Multimedia Applications. IEEE MultiMedia 14(1) (2007)
5. Lagoze, C., Hunter, J.: The ABC Ontology and Model. J. Digit. Inf. (2001)
6. Kiryakov, A., Popov, B., Kirilov, A., Manov, D., Ognyanoff, D., Goranov, M.: Semantic Annotation, Indexing, and Retrieval. In: 2nd International Semantic Web Conference, Florida, USA (2003)
7. Scherp, A., Franz, T., Saathoff, C., Staab, S.: F-a model of events based on the foundational ontology dolce+DnS ultralight. In: 5th International Conference on Knowledge Capture (K-CAP), California, USA (2009)

8. Carmagnola, F.: The five ws in user model interoperability. In: 5th International Workshop on Ubiquitous User Modeling, Gran Canaria, Spain (2008)
9. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: An online lexical database. International Journal of Lexicography 3(4), 235–312 (1990)
10. Timberlake, A.: Aspect, tense, mood. In: Shopen, T. (ed.) Language Typology and Syntactic Description 3, Grammatical Categories and the Lexicon (Language Typology and Syntactic Description), pp. 280–333. Cambridge University Press, Cambridge (2007)
11. Pustejovsky, J., Castano, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., et al.: TimeML: Robust Specification of Event and Temporal Expressions in Text. In: AAAI Spring Symposium on New Directions in Question Answering, Tilburg, Netherlands (2003)
12. Li, Z., Wang, B., Li, M., Ma, W.: A probabilistic model for retrospective news event detection. In: SIGIR, pp. 106–113 (2005)
13. Filatova, E., Hatzivassiloglou, V.: Event-based Extractive summarization. In: ACL, pp. 104–111 (2004)
14. Li, W., Wu, M., Lu, Q., Xu, W., Yuan, C.: Extractive Summarization using Inter- and Intra- Event Relevance. In: ACL (2006)
15. Liu, M., Li, W., Wu, M., Lu, Q.: Extractive Summarization Based on Event Term Clustering. In: ACL (2007)
16. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing? Int. J. Hum.-Comput. Stud. 907–928 (1995)
17. Wang, X., Mamadgi, S., Thekdi, A., Kelliher, A., Sundaram, H.: Eventory – An Event Based Media Repository. In: ICSC, pp. 95–104 (2007)
18. Brank, J., Grobelnik, M., Mladenic, D.: A survey of ontology evaluation techniques. In: 8th International Multi-Conference Information Society (IS 2005), pp. 166–170 (2005)
19. Wang, W., Zhao, D., Zhao, W.: Identification of topic sentence about key event in Chinese News. Acta Scientiarum Naturalium Universitatis Pekinensis 47(5), 789–796 (2011)
20. Wang, W., Zhao, D., Zou, L., Wang, D., Zheng, W.: Extracting 5W1H Event Semantic Elements from Chinese Online News. In: Chen, L., Tang, C., Yang, J., Gao, Y. (eds.) WAIM 2010. LNCS, vol. 6184, pp. 644–655. Springer, Heidelberg (2010)