Chinese News Event 5W1H Semantic Elements Extraction for Event Ontology Population

Wang Wei
(Supervised by Prof. Zhao Dongyan)
Institute of Computer Science & Technology, Peking University, Beijing, China
Department of Electronic Technology, Engineering University of CAPF, Xi'an, China
wjwangwei@pku.edu.cn

ABSTRACT

To relieve "*News Information Overload*", in this paper, we propose a novel approach of 5W1H (*who, what, whom, when, where, how*) event semantic elements extraction for Chinese news event knowledge base construction. The approach comprises a key event identification step, an event semantic elements extraction step and an event ontology population step. We first use a machine learning method to identify the key events from Chinese news stories. Then we extract event 5W1H elements by employing the combination of SRL, NER technique and rule-based method. At last we populate the extracted facts of news events to NOEM, an event ontology designed specifically for modeling semantic elements and relations of events. Our experiments on real online news data sets show the reasonability and feasibility of our approach.

Categories and Subject Descriptors

H.4.0 [**Information Systems Applications**]: General; I.7.2 [**Text Processing**]: Document—*languages*

General Terms

Documentation, Languages

Keywords

5W1H, event extraction, ontology, semantic role labeling

1. INTRODUCTION

The explosive growth of online news and users aggravate "*News Information Overload*". Classification, summarization and recommendation have been widely used to help people effectively access information. While existing technologies deal with documents based on "Bag of Words" model, which fails to provide sufficient semantic information about events. But people need more intelligent event-semantic level news services which can (1) push events but not documents to them and (2) provide entity and relation navigation among events to facilitate news browsing.

What people need, i.e., entities, relationships and events, can be extracted from text by using EE (Event Extraction) techniques. Considering the granularity of EE, existing work can be categorized into "atomic event" and "thematic event" extraction in sentence level or multi-document level, respectively. "Atomic event" extraction [2] [9] aims at extracting events in sentences by identifying actions (verbs or nominal verbs) and participants (e.g., person, location, time) of the event connected by the actions. "Thematic event" extraction, such as NEUXS [8], tries to extract several sub-events about a core event from clustered news articles and integrate them into an event framework to address the whole thing.

From this point of view, the state-of-the-art EE technology can not meet the requirement of event-semantic level news service. The granularity of atomic event is too small to be practical and the thematic event extraction is too coarse to include detailed and accurate event information. So extracting key event information from a single news story is a good choice. Due to the large scale and real-time update characteristics of news, although we only concern about the key events, it is sufficient to construct an event knowledge base.

In this paper, we discuss semantic understanding of Chinese news by extracting entities, relations involved in a key event of a news story. We adapt 5W1H (*what, who, when, where, why* and *how*), a concept in journalism, to represent semantic elements of news events, and propose a novel framework to address the whole list of 5W1H. Our approach comprises three steps: (1) A key event identification step which finds topic sentences that contain key event by measuring a sentence's importance on surface and semantic features of text according to style characteristics of news articles. (2) An event semantic elements extraction step which extracts 5W1H elements from the identified topic sentences using SRL (Semantic Role Labeling) and NER (Name Entity Recognition). (3) An event knowledge base construction step which first describes 5W1H elements with an ontology based event model NOEM (News Event Ontology Model) and then populates the ontology with the extracted event facts automatically.

The rest of the paper is organized as follows. Related work is reviewed in Sec. 2. The proposed methods are discussed in Sec. 3. Sec. 4 demonstrates evaluations and discussions. Sec. 5 concludes the paper and sketches our future work.

2. RELATED WORK

EE is a high-level IE (Information Extraction) task which tries to formulate an event as "*who* did *what* to *whom*, *when* and *where*". Formally, it automatically identifies events in free text and to derive detailed information such as time, location, participants and their roles in the events. It was primitively promoted by MUC (Message Understanding Con-

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

ferences) in 1987-1997 and then driven by ACE (Automatic Content Extraction) from 2000. A considerable amount of work as well as some profound thoughts on event extraction and synthesis have been reported [3].

MUC [5] takes EE as a domain-dependent scenario template filling task. The main research efforts focus on how to use lexical and syntax rules to match event patterns, and how to use unsupervised ML methods to get event extraction patterns automatically. NYU's Proteus[1] is a typical MUC EE system that built for several evaluations of topics (e.g. disaster, disease outbreak) in news domain.

An ACE event [1] involves zero or more ACE entities, values and time expressions. The goal of ACE VDR (Event Detection and Recognition) task is to identify all event instances, information about the attributes, and the event arguments of each instance of a pre-specified set of event types. David Ahn [2] break the task into a series of supervised machine learning sub-tasks to evaluate difficulty and importance of each task. Heng Ji [9] proposed a scheme of conducting cross-document inference to improve its result. However, due to small corpora and heavy linguistic technologies such as dependency parsers and NERs (Named-Entity Recognizers), precision/recall figures oscillating around 60% in these work are considered to be good results.

SRL is a task of identifying arguments for a predicate and assigning semantically meaningful labels to them. English SRL has achieved a good performance for practical EE tasks. Surdeanu [15] designed a domain-independent IE paradigm, which fills event template slots with predicate and their arguments identified automatically by a SRL parser. McCracken [12] used a SRL system to extract event from texts in a summary report genre.

Our work is a combination of ACE and SRL. We detect events which satisfy the ACE's definitions of event and event type/subtype. And by refining the result of SRL and NER, we extract event facts and map them to 5W1H elements in order to semantically understand an event. Our work is different from [12] in that we try to give a complete view on dealing with Chinese news story in online news domain.

3. EVENT SEMANTIC ELEMENTS EXTRACTION AND ONTOLOGY POPULATION

3.1 Methodology

The proposed framework for the semantic understanding of Chinese news is shown in Figure 1. We divide our approach into six sub-tasks and group them in three steps: (1) Title classification and topic sentences extraction for key event identification; (2) Semantic role labeling and 5W1H elements identification for event semantic elements extraction; (3) NOEM definition and Ontology population for event knowledge base construction.

Firstly, we identify key events of news stories. According to the rules in journalism writing, news stories have three characteristics: (1) One story mainly tells one thing; (2) Mostly, the headline of a news story contains essential information; (3) A topic sentence, which generally is present at the beginning of an article, tells key event of the news. Based on this observation, we extract topic sentences which contains the key events by using surface and semantic features of text and addressing the importance of news headline.

[1]http://nlp.cs.nyu.edu/index.shtml

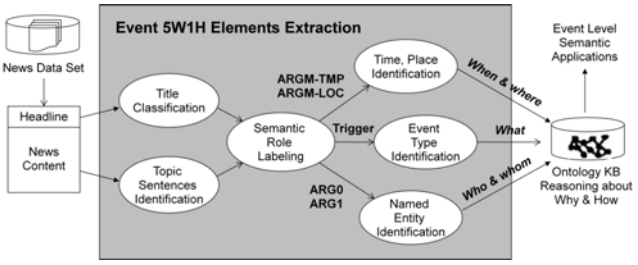

Figure 1: Chinese news event semantic understanding framework.

Secondly, we extract 5W1H semantic information from the key events. We employ a Chinese SRL parser to label semantic roles of constituents that have direct relationship with the predicates (event triggers) in a topic sentence. Typical semantic roles include agent, patient, source, goal and so on, which are core to a predicate, as well as location, time, manner, cause and so on, which are peripheral [18]. Such semantic information is important in answering 5W1H questions of a news event. We refine the results of the Chinese SRL parser by two methods: (1) using a list of trigger words to filter interesting events, and (2) using NER and heuristic rules as supplements.

At last, we employ SW (Semantic Web) techniques [4] to represent the extracted semantic information of events and build event knowledge base. Ontology is the core of SW and can be used to model concepts, relations of a domain. We design NOEM to represent 5W1H semantic elements of an event and relations among events. The extracted semantic event elements are first represented as RDF (Resource Description Framework) triples [2], and then they are automatically imported into NOEM as instances. By this means, event information is given well-defined meaning, enabling computers and people to work in co-operation.

3.2 Key Event Identification

A piece of news might describe more than one events, but only one is the key event. According to the general structure of news stories, the main information about the key event often appears in the headline or the first paragraph. So, by analyzing information content of a headline and other text features, a topic sentence which contains 5Ws about the key event can be identified. We give some definitions below.

DEFINITION 1. *News Event 5Ws: The {Time, Location, Subject, Predicate, Object} information which describe the* ***when, where, who, what, whom*** *of an event are called news event 5W elements. The 5W is denoted with the tuple <T,L,S,P,O>, where S, P, O are core elements and T, L are subordinates.*

DEFINITION 2. *Informative Headline: If a headline contains at least 1 element in <S,P,O> of a news event, in another word, if it can tell* ***who****,* ***what*** *or* ***whom*** *about the main event in a news story, the headline is informative.*

DEFINITION 3. *Topic Sentence: A sentence is a topic sentence iff the core elements <S,P,O> and at least one subordinate elements* ***T*** *or* ***L*** *can be extracted directly from the sentence.*

[2]RDF triples are used to group any two pieces of data and the link that connects them on the web.

3.2.1 Title Classification

Apparently, news title plays a key role to attract people in the first glance. Most of the time, titles attract readers with substantial information. But in some cases, they try to catch people's eye with exaggerated words which have nothing to do with the key events. A binary classification algorithm, which extracts keywords of a news story and counts the number of these keywords appeared in the title, is used to classify the title as informative or non-informative. We use Eq. (1) to calculate the similarity between the title H and the topic-word-set T of a news story.

$$Score_{ht} = \sum_{w \in H \cap T} 1 \tag{1}$$

The similarity score is determined by the number of overlap words between the title H and the topic-word-set T. When using *tfidf* or *PageRank* based methods, only notional words such as nouns, verbs and adjectives are selected for topic-word-set T. So, in all probability, topic words may appear in an informative title as core elements **S, P, O** to prompt an event. According to Def. 2, we set $Score_{ht} = 1$ as a threshold to identify the informativeness of a headline.

3.2.2 Topic Sentence Identification

For topic sentence identification, David Zajic employs a first-sentence selection method in headline generation task [7]. This is a coarse method because it is liable to be affected by datasets. According to the linguistic and structural characteristics of news style, we propose a method which use the surface and semantic features to pick up the most informative sentence. The features include term frequency, sentence location, sentence length, title words overlap rate, etc.

Term weight: Normalized term weight ($tfidf$) sum of notional words (no stop words) in a sentence.

$$Score_{term}(s_i) = \frac{\sum_{w \in s_i} term_{weight}(w)}{\max_{s_j \in d}(\sum_{w \in s_j} term_{weight}(w))} \tag{2}$$

Sentence Location: The 5W elements about the key event often appear at the beginning of a news story. Here we set the location *L=3*.

$$Score_{loc}(s_i) = \begin{cases} 1 & i \leq L \\ 1 - \frac{\log i}{\log n} & Otherwise \end{cases} \tag{3}$$

Sentence length: In general, a longer sentence contains more information about an event. Here we set sentence length *C=16*.

$$Score_{len}(s_i) = \begin{cases} 1 & if \quad Length(s_i) \geq C \\ 0 & Otherwise \end{cases} \tag{4}$$

Number of Name Entities: The 5W elements of a key event often appear as name entities such as time, place, person, organization and so on.

$$Score_{ne}(s_i) = \frac{\sum_{w \in s_i \cap \{NEs\}} 1}{Length(s_i)} \tag{5}$$

Title words overlap rate: The number of words that appear in both news headline and the sentence. We only concern about nouns, verbs and adjectives. This feature enables the implicit utilization of the semantic information in the headline and sentences.

$$Score_{hs}(s_i) = \frac{\sum_{w \in H \cap s_i} term_{weight}(w)}{\sum_{w \in H} term_{weight}(w)} \tag{6}$$

Informativeness of news headline: The number of NEs or proper nouns (actually the *who* and *whom* elements of the key event) that appear in the news headline.

$$Score_{hne} = \frac{\sum_{w \in H \cap \{NEs\}} 1}{Length(H)} \tag{7}$$

A linear combination of above features is used to measure the importance of each sentence. The score $SS(s_i)$ of sentence $s_i (i \leq n)$ is calculated as Eq. (8).

$$SS(s_i) = \sum_k w_k Score_k(s_i) \tag{8}$$

Where $k \in \{term, loc, len, ne, hs\}$ and w_k is the weight of each feature. Parameter w_k is tuned in our datasets. We use $Score_{hne}$ as the weight of $Score_{hs}$. For an informative headline, this will amplify the similarity between a sentence and the headline. For a non-informative headline, this will avoid its negative impact on the key event identification.

Sentences are ranked according to their $SS(s_i)$ scores and a threshold N is set for the number of selected sentences. So the Top-N ranked sentences are organized in a set, which is shorter than a summarization and more informative than the news headline. Taking the performance of the SRL parser into account, we believe that it is more simple and effective to deal with a few topic sentences instead of the whole text.

3.3 Event Semantic Element Extraction

The second step of our approach is to extract the 5W1H semantic elements for key event. We first use the HKUST Chinese Semantic Parser[3] to label semantic roles in the headline and topic sentences and then we improve the results with two methods: (1) using a list of trigger words to filter interesting events, and (2) using NER and heuristic rules to match semantic roles with 5W elements.

HKUST SRL parser tags all verbs as "TARGET" in the input sentence, and outputs related arguments, such as ARG0, ARG1, ARG2, ARGM-TMP and so on for each verb. The tagged verbs usually are event triggers, but some of them we are not interested in. So we extract 687 trigger words from ACE2005 Chinese training dataset, along with their event type-subtype information to build a trigger-event-type table. By querying trigger-event-type table, the key events can be easily recognized.

The result of the Chinese SRL parser is usually coarse and error-prone, then we use our self-implemented NER tool to identify Time, Place, Person and Organization entities for fine-granular event understanding. Thus we get 5W candidates of each event, i.e. predicate, event type, NEs, time and location words. After that, we use rule-based methods to map these candidates to the 5W1H semantic elements of an event. The detailed information about the implementation of the algorithm can be found in our work [16].

3.4 Ontology Definition and Population

3.4.1 News Event Ontology Model

Event modeling involves event definition, event information representation and storage. There are a number of event models in different domains nowadays. For example, the event templates used in MUC, a structural event representation in ACE, a generic event model E [17] in event-centric multimedia data management, and ontology based

[3]http://hlt030.cse.ust.hk/research/c-assert/.

event models such as ABC [11], PROTON [10] in knowledge management, EO (Event Ontology) [13] describing music production process and Event-Model-F [14] in distributed event-based system. But these models are not suitable for semantic understanding of Chinese news stories. In order to provide a basic vocabulary for semantic annotations with respect to the 5W1H of news stories, based on above models, we propose a News Ontology Event Model (NOEM) to describe entities and their relations in news events.

In accordance with Jain's generic event model, we design NOEM to capture the temporal, spatial, information, experiential, structural and causal aspects of events. The main concepts and properties of NOEM are shown in Figure 2. They cover three types of information about events: event elements, event relations and event media. A special property of NOEM is that we import ACE event hierarchy to identify event's types by trigger words, as well as CNML (Chinese News Markup Language) categories to represent a news article's topic so that we can connect an event to its category in document-level.

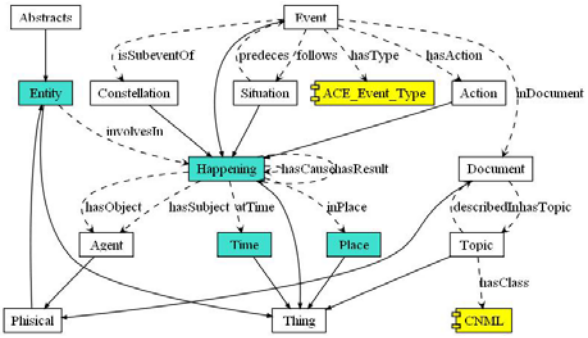

Figure 2: 5W1H News Event Ontology Model.

3.4.2 *Ontology Population*

NOEM is constructed in Protégé[4]. By using a predefined template, we can automatically generate an OWL (Web Ontology Language)[5] file for extracted event facts and import them into the ontology to build an event knowledge base.

For example, from sentence "[**Chinese President {Hu Jintao**}$_P$]$_{ARG0}$ [**arrived**]$_{TARGET}$ **in** [{**Ottawa**}$_L$]$_{ARG1}$, **capital of {Canada**}$_L$, **for a state visit on** [{**September 8**}$_T$]$_{ARGM-TMP}$", we can identify an ACE "Movement" event and map the extracted event 5w elements to concepts and relations in NOEM by triples <*Hu Jintao, arrive, Ottawa*>, <*arrive, isTypeof, Movement/Transport*>, <*Hu Jintao, isTypeof, Person*>, <*September 8, isTypeof, Time*> and <*Ottawa, isTypeof, Place*>. By populating triples into NOEM, an event knowledge base which comprises semantic relations among persons, time, places, topics and events will be built.

4. EVALUATION AND DISCUSSION

4.1 Data Set

To our knowledge, there are no suitable open Chinese news dataset for open evaluation. So we select some news stories and manually annotate them with title type, topic sentences and key event 5W elements to consturct three datasets: (1)ACE05. It contains 182 XinHua newswire articles which come from ACE2005 Chinese training corpus. (2) BJRB. It contains 543 news stories selected from March 1-10, 2009 Beijing Daily. (3) XAWB. It contains 808 news stories selected from June 1-10, 2009 Xi'an Evening News. Statistics of the three datasets are shown in Table1.

Table 1: Statistics of our three datasets.

Datasets	ACE05	BJRB	XAWB
Articles before selection	194	760	1279
Articles after selection	182	543	808
Average words in sentence	8.79	11.43	16.14
Average words in story	447	593	783
Number of informative title	182	512	741
First sentence is topic sentence	168	396	516

4.2 Key Event Identification

4.2.1 *Title Classification*

According to Eq. (1), we use $tfidf$ and $PageRank$ based methods to extract topic words from a news story and identify whether the news headline is informative. The results on the three datasets are shown in Table 2.

Table 2: Result of the title classification task.

Method	ACE05	BJRB	XAWB
TFIDF	0.994	0.966	0.922
PageRank	0.961	0.864	0.852

The evaluation result shows that the $tfidf$-based method has a better classification accuracy than the $PageRank$-based method. We believe the reason lies in that $PageRank$-based method can not identify some uncommon words (abbreviations or special words) in the title as topic words.

4.2.2 *Comparison of Feature Selection Strategy*

As shown in Eq. (8), we apply a three-fold cross-validation method to tune the parameters and set 0.1, 0.5, 0.1, 0.1 and 0.2 for feature weight *term, loc, len, ne* and *hs* respectively to identify a topic sentence. We use 6 types of feature combination strategies to weight sentences and choose the top score one as topic sentence. The results of different feature combinations are shown in Table 3.

Table 3: Comparison of algorithm precision with different feature combination.

System	ACE05	BJRB	XAWB
FS: first sentence	0.923	0.729	0.639
TS: only title similarity	0.868	0.646	0.580
NT: no title feature	0.626	0.527	0.298
WT: title not classified	0.907	0.716	0.625
TC: title classified	0.929	0.772	0.658
TG: title classified manually	0.929	0.772	0.661

The result shows that title and title classification are both important to find a topic sentence. $FS > TS > NT$ means that the first sentence is more important than the title, and the title is above other features. $FS > WT > TS$ means other features (*term, loc, len, ne*) are useful and helpful when combined with the title. $TC > FS > WT$ means title classification is very useful to mitigate the effect of non-informative titles. TC is as good as TG (the gold standard) proves our title classification method is effective.

[4]http://protege.stanford.edu/
[5]http://www.w3.org/TR/owl-ref/

4.2.3 Topic Sentences Identification

We test the topic sentence identification algorithm on all three datasets. If the extracted topic sentences contain the human tagged topic sentence, we mark the identification as correct, otherwise as wrong. We only use precision[6] as metric to evaluate the performance of our algorithm.

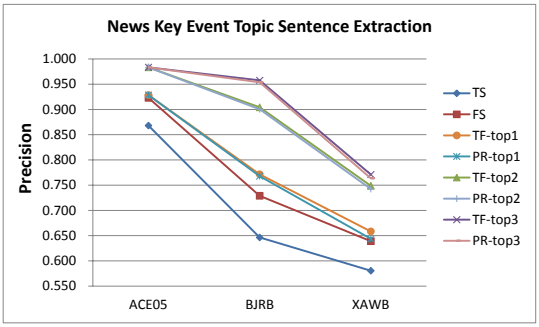

Figure 3: Result of the news key event topic sentence identification.

In Figure 3, we compare the accuracy of the proposed topic sentence identification algorithm to the baselines (FS and TS) when extracting top-N ($N = 1, 2, 3$) sentences, respectively. Our algorithm is better than the baselines when selecting one topic sentence. If the top-2 or top-3 sentences are extracted, the accuracy of our algorithm increased obviously. This means our algorithm has located the topic sentence in a small sentence set, but it is difficult to rank the most suitable one as top-1 because there are complicate influences among features. The result also shows that our algorithm tends to be affected by the document length.

4.3 5Ws Extraction

For the 5W-tuple evaluation, checking whether two tuples are identical only would penalize too much those whose tuples are almost correct. As in [6], we evaluate 5W-tuple by using a string similarity measure to compute the similarity between the extracted *Time*, *Location*, *Subject*, *Predictive*, *Object* elements of the 5W-tuple $< T, L, S, P, O >$ and manually annotated results.

Primitive evaluation results show "when" and "where" are better than "who" and "whom". It makes sense because time and location are comparatively easier to identify. What is more, "who" and "whom" are closely related to "what", so the accumulated errors lead to a lower precision.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel event semantic understanding framework to facilitate online news browsing. The significance of this approach is that it first identifies topic sentences of news stories by emphasizing the importance of the headline, and that it integrates SRL and NER to extract event 5W1H semantic elements for knowledge base construction. Experiments are conducted on each part of the pipeline on manually labeled real world datasets. The experimental results show the feasibility of our approach.

Our future work should be improving the performance of the proposed methods and the precision of the event 5W1H elements extraction algorithm. We aim at building a practical event extraction system and an event knowledge base to support event level semantic applications.

[6] $Precision = \frac{number\ of\ right\ extracted\ topic\ sentences}{number\ of\ news\ articles\ in\ dataset}$

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] ACE. *Chinese Annotation Guidelines for Events*. National Institute of Standards and Technology, 2005.

[2] D. Ahn. The stages of event extraction. In *Proceedings of COLING/ACL 2006 Workshop on Annotating and Reasoning about Time and Events*, 2006.

[3] N. Ashish, D. Appelt, D. Freitag, and D. Zelenko. Proceedings of aaai-06 workshop on event extraction and synthesis. Boston, Massachusetts, USA, 2006.

[4] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.

[5] N. Chinchor and E. Marsh. Muc-7 information extraction task definition (version 5.1). 1998.

[6] L. Dali and B. Fortuna. Triplet extraction from sentences using svm. In *Proceedings of SiKDD*.

[7] B. Dorr, D. Zajic, and R. Schwartz. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop*, 2003.

[8] P. Jakub, H. Tanev, and P. O. Wennerberg. Extracting Violent Events from On-line News for Ontology Population. In *Proceedings of the 10th International Conference on Business Information Systems (BIS'2007)*, pages 287–300, 2007.

[9] H. Ji and R. Grishman. Refining event extraction through unsupervised cross-document inference. In *Proceedings of the 46th ACL*, 2008.

[10] A. Kiryakov, B. Popov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. Semantic annotation, indexing, and retrieval. *J. Web Sem.*, 2004.

[11] C. Lagoze and J. Hunter. The abc ontology and model. *Journal of Digital Information*, 2001.

[12] N. McCracken, N. Ozgencil, and S. Symonenko. Combining techniques for event extraction in summary reports. In *Proceedings AAAI06 Workshop on Event Extraction and Synthesis*, 2006.

[13] Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson. The music ontology. In *the International Conference on Music Information Retrieval*, pages 417–422, 2007.

[14] A. Scherp, T. Franz, C. Saathoff, and S. Staab. F-a model of events based on the foundational ontology dolce+dns ultralight. In *Conference on knowledge capturing (K-CAP)*, 2009.

[15] M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. Using predicate-argument structures for information extraction. In *Proceedings of the 41st ACL*, pages 8–15. ACL, 2003.

[16] W. Wang, D. Zhao, and D. Wang. Chinese news event 5w1h elements extraction using semantic role labeling. In *Proceedings of the third International Symposium on Information Processing*, pages 484–489, 2010.

[17] U. Westermann and R. Jain. Towards a common event model for multimedia applications. *IEEE MultiMedia*, pages 19–29, 2007.

[18] N. Xue. Labeling chinese predicates with semantic roles. *Computational Linguistics*, 34:225–255, 2008.