Extracting 5W1H Event Semantic Elements from Chinese Online News*

Wei Wang[1,2], Dongyan Zhao[1,3], Lei Zou[1], Dong Wang[1], and Weiguo Zheng[1]

[1]Institute of Computer Science & Technology, Peking University, Beijing, China
[2]Engineering College of Armed Police of People's Republic of China, Xi'an, China
[3]Key Laboratory of Computational Linguistics (Peking University),
Ministry of Education, China
{wangwei,zdy,zoulei,wangdong,zhengweiguo}@icst.pku.edu.cn

Abstract. This paper proposes a verb-driven approach to extract 5W1H (**W**ho, **W**hat, **W**hom, **W**hen, **W**here and **H**ow) event semantic information from Chinese online news. The main contributions of our work are two-fold: First, given the usual structure of a news story, we propose a novel algorithm to extract topic sentences by stressing the importance of news headline; Second, we extract event facts (i.e. 5W1H) from these topic sentences by applying a rule-based method (verb-driven) and a supervised machine-learning method (SVM). This method significantly improves the predicate-argument structure used in Automatic Content Extraction (ACE) Event Extraction (EE) task by considering valency (dominant capacity to noun phrases) of a Chinese verb. Extensive experiments on ACE 2005 datasets confirm its effectiveness and it also shows a very high scalability, since we only consider the topic sentences and surface text features. Based on this method, we build a prototype system named Chinese News Fact Extractor (CNFE). CNFE is evaluated on a real world corpus containing 30,000 newspaper documents. Experiment results show that CNFE can extract event facts efficiently.

Keywords: Relationship Extraction, Event Extraction, Verb-driven.

1 Introduction

To relieve information overload, some techniques have been proposed based on Machine Learning (ML) and Natural Language Processing (NLP), such as classification, summarization, recommendation and Information Extraction (IE). These techniques are also used in online news browsing to reduce the pressure of explosive growth of news articles. However, they fail to provide sufficient semantic information for event understanding since they only focus on the document instead of the event itself. In order to get more semantic information about the event, some event-oriented techniques have been proposed, such as Event-based summarization [1,2,3] , Topic Detection and Tracking (TDT) and

* This work is sponsored by Beijing Municipal Science & Technology Commission project "R & D of 3-dimensional risk warning and integrated prevention technology" and China Postdoctoral Science Foundation (No.20080440260).

L. Chen et al. (Eds.): WAIM 2010, LNCS 6184, pp. 644–655, 2010.
© Springer-Verlag Berlin Heidelberg 2010

so on. However, these event-oriented techniques consider semantic information in a coarse-grained manner.

In this paper, we discuss how to extract structural semantic information from online news corpus, which is the first step to build a large-scale online news semantic knowledge base. In order to address this issue, we consider an important concept in news information gathering, that is 5W1H. The 5W1H originally states that a news story should be considered as complete if it answers a checklist of six questions, what, why, who, when, where, how. The factual answers to these six are considered to be elaborate enough for people to understand the whole story [4]. Recently, 5W1H concept is also utilized in Event Extraction (EE) and Semantic Role Labeling (SRL). However, due to "heavy" linguistic technologies such as dependency parsers and Named-Entity Recognizers (NERs), it is computationally intractable for EE and SRL over a large news corpora [5]. Even though we can use the grid infrastructure, the computational cost is still too high. Therefore, in this paper, we propose a "light" but effective method to extract 5W1H facts from a large news corpora.

We propose a novel news event semantic extracting method to address 5W1H. This method includes three steps: topic sentences extraction, event classification and 5W elements extraction. First, given the importance of news headline, we identify informative sentences which contain the main event's key semantic information in the news article. Second, we combine a rule-based method (verb-driven) and a supervised machine-learning method (SVM) to extract events from these topic sentences. Finally, we recognize 5Ws with the help of specific event templates as well as event trigger's valency and syntactic-semantic rules. We treat the topic sentences, actually a short summarization of the news as "How" of the event and replace "Why" with "Whom" currently. Thus, we obtain a tuple of 5W <Time, Location, Subject, Predicate, Object> and "How".

Based on the proposed method, we implement CNFE (Chinese News Fact Extractor). We evaluate CNFE on a real world corpus containing more than 30,000 newspaper documents to extract ACE events. Experiment results show that CNFE can extract high-quality event facts (i.e. 5W1H) efficiently.

To summarize, we made the following contributions in this paper:

- We propose using 5W1H concept to formulate an event as "[Who] did [What] to [Whom], [When], [Where] and [How]", as these information are essential for people to understand the whole story.
- To extract 5W1H efficiently, given structural characteristics of news stories, we propose a novel algorithm to identify topic sentences from news stories by stressing the importance of headline.
- We propose a novel method to extract events by combining a rule-based (verb-driven) method and a machine-learning method (SVM). The former considers valency of an event trigger. To the best of our knowledge, we are the first to introduce valency grammar into Chinese EE.
- We perform extensive experiments on ACE2005 datasets and a real news corpus, and also compare CNFE with existing methods. Experiment results confirm both effectiveness and efficiency of our approach.

The rest of the paper is organized as follows. In section 2, we discuss some related works in details. The whole flow of proposed event 5W1H semantic elements extraction approach is discussed in section 3. Sequentially in section 4, we demonstrate the results of our experiments on our methods and the prototype system CNFE. Finally, we draw a conclusion in section 5.

2 Related Works

IE refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources. Primitively promoted by MUC in 1987-1997 and then developed by ACE since 2000, studies in IE have shifted from NE recognition to binary and multi-way relation extraction. Many research works focus on triple extraction for knowledge base construction, such as Snowball [6], Knowitall [7], Textrunner [5], Leila [8] and StatSnowball [9]. These works confirm the importance of technologies such as pattern matching, light natural-language parsing and feature-based machine learning for large scale practical systems.

Event extraction is actually a multi-way relationship extraction. In MUC-7 [10], event extraction is defined as a domain-dependent scenario template filling task. An ACE [11] event is an event involving zero or more ACE entities, values and time expressions. The goal of the ACE Event Detection and Recognition (VDR) 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. ACE defines the following terminologies related with VDR:

- Event: a specific occurrence involving participants. An ACE event has six attributes (type, subtype, modality, polarity, genericity and tense), zero or more event arguments, and a cluster of event mentions.
- Event trigger: the word that most clearly expresses an event's occurrence.
- Event argument: an entity, or a temporal expression or a value that has a certain role (e.g., Time-Within, Place) in an event.
- Event mention: a sentence (or a text span extent) that mentions an event, including a distinguished trigger and involving arguments.

Driven by ACE VDR task, Heng Ji proposed a serial of schemes on event coreference resolution [12], cross-document [13,14] and cross-lingual [15] event extraction and tracking, these schemes obtained encouraging results. David Ahn decomposed VDR into a series of machine learning sub-tasks (detection of event anchors, assignment of an array of attributes, identification of arguments and assignment of roles, and determination of event coreference) in [16]. Results show that arguments identification has the greatest impact of about 35-40% on ACE value and trigger identification has a high impact of about 20%. Naughton investigated sentence-level statistical techniques for event classification in [17]. The results indicate that SVM consistently outperform the Language Model (LM) technique. An important discovery is that a manual trigger-based classification approach (using WordNet to manually create a list of terms that are synonyms or hyponyms for each event type) is very powerful and outperforms the SVM on three of six event types. However, these works mainly focus on English articles.

Chinese information extraction research starts relatively late, the main research work focused on the Chinese Named Entity Recognition, as well as the mutual relations between these entities. The latest development of Chinese event extraction in the past two years was reported in [18] and [19]. In [18], Yanyan Zhao et. al. proposed a method combining event trigger expansion and a binary classifier in event type recognition and a one-with-multi classification based on Maximum Entropy (ME) in argument recognition. They evaluated the system on ACE2005 corpus and achieved a better performance in comparison with [16]. However, their work left the problem of polysemy unsolved, i.e. some verbs can trigger several ACE events so that they might cause misclassification.

SRL is another example of multi-way semantic relation extraction. Different from full semantic parsing, SRL only labels semantic roles of constituents that have direct relationship with the predicates (verbs) in a 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 [20]. Such semantic information is important in answering 5W1H of a news event. Surdeanu [21] designed a domain-independent IE paradigm, which filled event template slots with predicate and their arguments identified automatically by a SRL parser. However, semantic parsing is still computationally intractable for larger corpus.

Aiming at building a scalable practical system for Chinese online news browsing, we propose a method combining a verb-driven method and a SVM rectifier to identify news event. We also use trigger's valence information and syntactic-semantic rules to help extracting event facts. We keep our research consistent with ACE event extraction in order to compare with other works.

3 Event 5W1H Elements Extraction

3.1 Preliminary

Before the formal discussion of our method, we first give our observations about online news. After observing more than 6000 Chinese news stories in two famous online news services, xinhuanet.cn and people.com.cn, we find that online news stories have three special characteristics: 1) One news story usually tells one important event; 2) Being an eye-catcher, headline often reveals key event information. Furthermore, it contains at least two essential elements such as "*W*ho" and "*W*hat"; 3) Usually, in the first or second paragraph of the story, there is a topic sentence which expands the headline and tells the details of the key event. According to our statistics, the percentages are 74.6% and 9.8% respectively. A feasibility research in [22] also reveals the importance of headlines and topic sentences. Actually, all the characteristics precisely agree with news articles' writing rules. We base our idea of topic sentences extraction on these observations.

We utilize Chinese valency grammar in our 5W1H extraction method. Valency grammar [23,24] was first proposed by French linguist Lucien Tesniere in 1953 and was led into Chinese grammar by Zhu Dexi in 1978. Modern Chinese valency grammar tries to characterize the mapping between semantic predicate-argument relationships and the surface word and phrase configurations by which they are expressed. In a predication which describes an event or an action caused by a verb, arguments (participants of the event) play different roles. Some common accepted semantic roles are agent, patient, dative, instrument, location, time, range, goal, manner and cause. Among them, agent, patient and dative are obligatory arguments, while others are optional. The number of obligatory arguments is valence of a verb, prompting the dominant capacity of a verb to NPs. Research on the valence grammar has lasted for more than 20 years in China, and many progresses have been made. Although the research scope has been extended from verbs to adjectives and nouns, as a feasibility research, verbs are our only concern in this paper. Two example sentences containing a bivalent and a trivalent verb respectively are given below. They demonstrate that a polyseme may have different valencies for different meanings.

- In sentence "我送小王去火车站。(I walk Xiao Wang to the railway station.)", "walk" is a bivalent verb, "I" is subject (agent), "Xiao Wang" is object (patient) and "to the railway station" is a complement (goal). The sentence complies to "NP1+V+NP2" and "walk" dominates two NPs.
- In sentence "小王送我一本书。(Xiao Wang gives me a book.)", "Xiao Wang" is agent, "me" is dative and "a book" is patient. They are obligatory arguments of verb "give". The sentence is not complete if absence any one of them. So valency of "give" is 3. This kind of sentences comply to the syntactic structure "NP1+V+NP2+NP3" or "NP1+PNP2+V+NP3".

According to [25], most verbs in Chinese are bivalent, there are only about 573 trivalent verbs. In [26], a classification and distribution of 2056 meanings of 1223 common verbs in Chinese Verbs Usage Dictionary is investigated. It shows that there are 236 univalent verbs (11.5%), 1641 bivalent verbs (79.8%) and 179 trivalent verbs (8.7%). In order to extract obligatory arguments of an event, we collect syntactic patterns of verbs with different valencies. We use regular expressions coding these syntactic-semantic rules and match them in a sentence to improve the precision of EE. The syntactic-semantic patterns and rules for distinguishing different types of verbs are shown in Table 1.

Table 1. Classification and syntactic-semantic rules of verbal valency

Verb Type	Examples	Rules	Syntactic Patterns
univalent	来,地震,游泳	$(a \vee b) \wedge (\sim (c \wedge d \wedge e \wedge f))$	a: NP1+V, b: V+NP1, c: NPl+V+NP2, d: NP1+PNP2+V, e: NP1+V+NP2+NP3, f: NP1+PNP2+V+NP3
bivalent	泼,喜欢,挥手	$(c \vee d) \wedge (\sim (e \wedge f))$	
trivalent	打听,供给,存放	$e \vee f$	

3.2 Algorithm Overview

Based on the three assumptions and Chinese valency grammar, we defined the 5W1H elements extraction task as: given a headline of a news article, which often contains at least 2Ws (Who & What), find a topic sentence containing the most important event of the story, and extract other Ws from the topic sentences for specified events. The output is 5W1H: a tuple of <Time, Location, Subject, Predicate, Object> and "How" of the event.

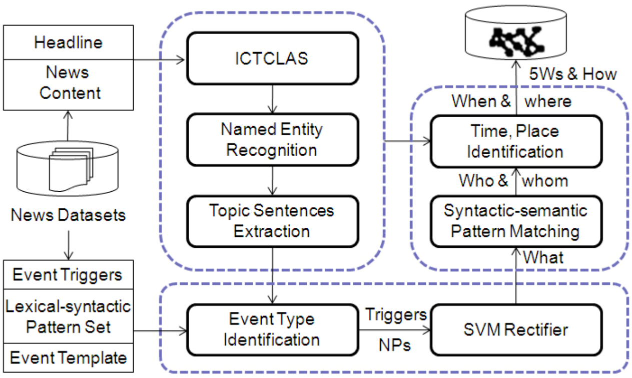

Fig. 1. The framework of event 5W1H semantic elements extraction

Fig. 1 shows the framework of the proposed method. The method consists of three main phases: topic sentence extraction, event type identification and 5W extraction. First, for a given news article, after word segmentation and POS tagging using ICTCLAS[1], we design a novel algorithm to extract topic sentences by stressing the importance of news headlines. Second, we identifies events and their types with the help of a list of trigger words extracted from ACE training corpus and further refines the results using a SVM classifier. In the third phase, we recognize event arguments according to the trigger's valency and its syntactic-semantic rules. At last we output 5W tuples and use the ordered topic sentences as a short summarization to describe "How" of the event.

3.3 Topic Sentences Extraction

Topic sentence identification task is to find an informative sentence which contains key information (5Ws) of a news story. In David Zajic's work of headline generation [22], he directly chose the first sentence as topic sentence based on his observation that a large percentage of headline words are often found in the first sentence. Unfortunately it is not always the case. So we employ extractive automatic text summarization technology to find salient sentences using surface-level features in this work.

According to the linguistic and structural features of a news story, that is, trying to attract readers' eyeballs with a headline or at the beginning of the text, we consider term frequency [27], sentence location in text [28], sentence length [29] and title words overlap rate [30] to select topic sentences. Equation (1) shows the proposed function for sentence importance calculation.

[1] Institute of Computing Technology, Chinese Lexical Analysis System. http://www.ictclas.org/index.html

$$SS_i = \alpha \frac{f(\sum_{w \in S_i} tf(w) \cdot idf(w))}{Length_i} + \beta f(Length_i)$$
$$+\gamma f(Position_i) + \delta \frac{f(\sum_{w \in S_i} \sum_{w \in T} 1)}{|T| \cdot Length_i} \quad (1)$$

Where α, β, γ, and δ are the parameters (positive integers) for term frequency ($tfidf$), sentence length, sentence location and title's weights, respectively. $f(x)$ is normalization for each parameter. $f(x) = x/(\sum_{S_i \in C} x_i)$

Sentences are ranked by SS_i score. By setting a proper threshold N for the number of selected sentences, we choose the N-Best as the topic sentences and get a sentence set, which is more informative than news headline and shorter than normal summarization. Taking the performance and scalability into account, we believe that it is more simple and effective to deal with a few topic sentences instead of the whole text.

3.4 Event Type Identification

The event type identification module accepts three inputs as shown in Fig. 1: topic sentences with word segmentation and POS tags, a trigger-event-type/subtype table and a set of syntactic-semantic rules of triggers. The list of triggers and their event type/subtype are extracted from ACE05 training dataset. Additionally, event templates of each type/subtype are associated to the triggers. The valency information of each trigger's different meanings and corresponding syntactic-semantic patterns are built based on [25] and [26].

The type identification module searches the topic sentences by examining trigger list and marks each appearance of a trigger as a candidate event. The module also finds chunks using some heuristic rules: 1) Connecting numerals, quantifiers, pronouns, particles with nouns to get max-length NPs; 2) Connecting adjacent characters with same POS tags; 3) Identifying special syntactic patterns in Chinses such as "de(的)", "bei(被)", "ba(把)" structures. After that, the identified trigger, candidate event type/subtype, original sentences with POS tags, NEs and newly identified chunks are input to a SVM rectifier to do fine-grained event type identification. The SVM rectifier is trained on ACE05 Chinese training dataset with deliberately selected features so that it can find out the wrong classification of trigger-based method. By employing lexical and semantic information of a trigger, it can partly solve the polysemy problem and improve the precision of event classification. Features used in the SVM rectifier include:

Trigger's Information: Numbers of a polysemous trigger's meanings, the trigger's frequency on one type of event normalized by its total event frequency, the trigger's total event frequency normalized by its word frequency.

Trigger's NE Context: Presence/absence of two NEs before and behind the trigger with type of Person, Organization and Location.

Trigger's Lexical Context: Presence/absence of POS tags (N, V, A, P and others) of two words before and behind the trigger.

Other Features: Sentence length, the presence/absence of a time-stamp and a location term.

3.5 5W1H Extraction

From the output of event classification, we get headline, topic sentences and a list of 5W candidates of an event, i.e. predicate, event type, NEs, time and location words. Next is to identify semantic elements for the event.

What: Based on our assumptions, we believe that title and topic sentences contain the key event of the news. So we use the event type which is first identified by verb-driven method and then rectified by SVM as "What".

Who, Whom: To identify these two arguments, we analyze the topic sentences in syntactic and semantic planes. In syntactic plane, the NPs and special syntactic structures such as "de(的)", "bei(被)", "ba(把)" are found. In semantic plane, regular expressions are used to match trigger's syntactic-semantic rules. For example, we use an expression "(.*)/n(.*)/trigger(.*?)/n(.*?)/n.*" to match "NP1+V+NP2+NP3". The rules are downward-compatible, i.e., trivalent verbs can satisfy bivalent verb's patterns but not vice versa. So we examine the patterns from trivalent verbs to univalent verbs. We identify obligatory arguments from NEs and NPs of the trigger according to the sentence's syntactic structures. Then we determine their roles (e.g. agent, patient) and associate them with a specific event template.

When, Where: We combine outputs of NER and ICTCLAS to identify time and location arguments. Priority is given to NER. If there is no Time/Location NEs, generated chunks with tags of /nt and /ns are adopted.

How: We use contents of identified <Subject, Predicate, Object> as "How" to describe the process of an event. We first order the triples by examining where they appear in the topic sentences, and then extract the contents between subject and object to describe "Who did What to Whom".

4 Evaluation and Discussion

To evaluate our method, we conduct three experiments. The first experiment is conducted on a self-constructed dataset to evaluate the topic sentence extraction method. The second experiment is conducted on a benchmark dataset to evaluate the effectiveness of our 5W element extraction algorithm. The third experiment is conducted on an open dataset to measure the scalability of CNFE.

4.1 Data Set

DS1 is a self-constructed dataset comprises of 235 Chinese news stories in ACE 2005 training dataset's news wire section and 765 latest news stories collected from xinhuanet.cn. For each story, a topic sentence is labeled manually.

DS2 is the ACE 2005 Chinese training corpus. We extract labeled events from it and there are 2519 sentences in total. These sentences are used to evaluate the verb-driven 5W1H extracter and the SVM rectifier.

DS3 is a whole collection of Beijing Daily's online news in 2009, which contains 30,000 stories. We use this corpus to evaluate the performance of CNFE.

4.2 Evaluation of Topic Sentences Identification

Topic-sentence extraction is the basis of the 5W1H extraction. To identify a topic sentence, as described in equation (1), we consider a sentence's word frequency ($tfidf$), its length, location and word co-occurrence with the headline. To emphasize the weight of title, we fix parameters α, β and γ, and variate parameter δ. We evaluate the precision of topic sentences extraction on DS1 manually. At first we only extract the top scored sentence, but we find some Ws are lost. So we lower the threshold to top three sentences so that more details about the event can be extracted. If the extracted top 3 topic sentences contain the human tagged topic sentence, we mark an extraction as true, otherwise as false.

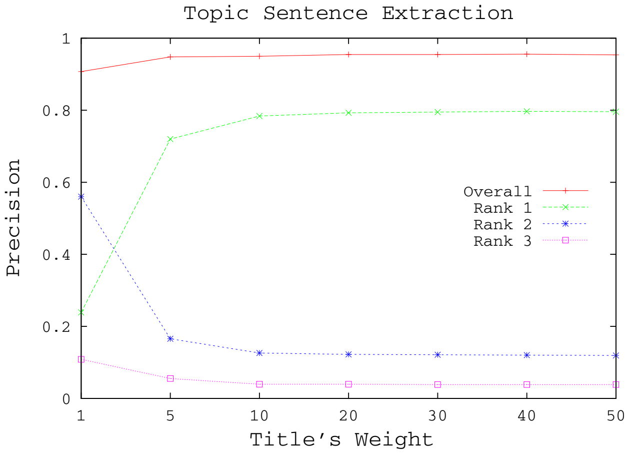

Fig. 2. The precision of topic sentence extraction along with title's weight

Fig. 2 shows the result of topic sentences extraction. With title's weight increases gradually, the precision of topic sentence identification gets a higher value. We choose 40 for parameter δ and achieve a precision of 95.56%. Apparently, the precision of topic sentences extraction on our data set is promising.

4.3 Evaluation of 5W Extraction

In this section, we evaluate our event detection algorithm and the quality of extracted 5W tuples. Our event detection and classification method is verb-driven and enhanced by SVM, as described in 3.4. We extract 621 triggers (verbs) from DS2 and find that only 48 of them are multi-type event triggers. By querying trigger-event-type table, we get candidate event types for each trigger appears in the topic sentences. When a trigger is a polyseme or tagged as a noun, the verb-driven method will not work. Then the SVM classifier, which is implemented on LibSVM[2] and trained on DS2 with deliberately selected features, can make a decision. For evaluation, we use classic F1 score and compare it with [18], which used a similar method. The result is shown in Table 2. For 5W-tuple evaluation, a way

Table 2. Results for event detection and classification

Methods		Recall	Precision	F1
Proposed methods	Verb-driven	74.43%	48.35%	58.62%
	Verb-driven+SVM	68.31%	57.27%	**62.30%**
Yanyan Zhao	trigger expansion+binary classifier	57.14%	64.22%	60.48%

of comparing tuples is needed. To just check whether two tuples are identical would penalize too much those which tuples are almost correct. We extend the idea of triple evaluation defined in [31] to 5W-tuple evaluation, which employs a string similarity measure to compute the similarity between extracted T, L, S, P, O elements and annotated results in DS2. We carefully examined the tuples extracted from "Movement" events (633 event mentions) and "Personnel" events (199 event mentions) from DS2 by hand to find problems of our method. The result is shown in Table 3. We find that the most important factor that affects the correctness of 5W is the complexity of language. Compound sentences and special syntactic structures of Chinese make our extractor error-prone. Wrong segmentation and POS tags, for example, a trigger is segmented into two words and a verb trigger is wrongly tagged as a noun, have a strong impact on the result. The main problems of our method which cause wrong assignment of arguments lie in absence of coreference resolution and wrongly identified NPs.

Table 3. Results of extracted 5Ws on Movement and Personnel events in DS2

ACE Event	Right		Wrong		
(number)	T, L, S, P, O	$T, L, P, (S\|O)$	POS	Structure	Method
Movement(633)	240	161	65	96	71
Personnel(199)	99	25	26	28	21

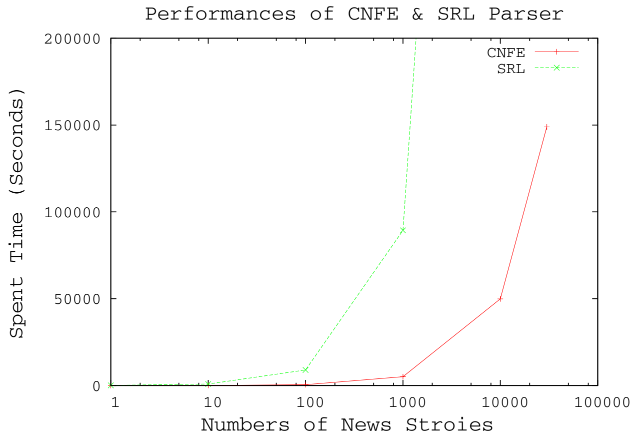

Fig. 3. The performances of CNFE and the SRL baseline system

[2] URL: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

4.4 Evaluation of CNFE

Sine CNFE only use surface text features to extract event facts, it has a better scalability than SRL parser. We implement a baseline system to tag semantic predicate argument structure based on HKUST Chinese Semantic Parser[3]. We run the two systems on DS3 and the performances of them are shown in Fig. 3.

5 Conclusions

In this paper, we propose a novel method to extract 5W1H event semantic information from Chinese online news. We make two main contributions in our work: First, based on a statistic analysis of structural characteristics of 6000 news stories, we propose a novel algorithm to extract topic sentences from news stories by stressing the importance of headline. Second, we propose a method of combining a rule-based method (verb-driven) and a supervised machine-learning method (SVM) to extract 5W1H facts from topic sentences. This method improves predicate-argument structure used in ACE EE task by considering valency of Chinese verbs. Finally, we conduct extensive experiments on a benchmark dataset and an open dataset to confirm the effectiveness of our approach.

Acknowledgments

We would like to thank Sameer Pradhan and Zhaojun WU for offering HKUST Chinese Semantic Parser.

References

1. Filatova, E., Hatzivassiloglou, V.: Event-based Extractive summarization. In: Proceedings of ACL, pp. 104–111 (2004)
2. Li, W., Wu, M., Lu, Q., Xu, W., Yuan, C.: Extractive Summarization using Inter- and Intra- Event Relevance. In: Proceedings of ACL (2006)
3. Liu, M., Li, W., Wu, M., Lu, Q.: Extractive Summarization Based on Event Term Clustering. In: Proceedings of ACL (2007)
4. Carmagnola, F.: The five ws in user model interoperability. In: UbiqUM (2008)
5. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open Information Extraction from the Web. In: Proceedings of IJCAI, 2670–2676 (2007)
6. Agichtein, E., Gravano, L., Pavel, J., Sokolova, V., Voskoboynik, A.: Snowball: A Prototype System for Extracting Relations from Large Text Collections. In: Proceedings of SIGMOD Conference, pp. 612–612 (2001)
7. Etzioni, O., Cafarella, M.J., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Web-scale information extraction in knowitall (preliminary results). In: Proceedings of WWW, pp. 100–110 (2004)
8. Suchanek, F.M., Ifrim, G., Weikum, G.: Combining linguistic and statistical analysis to extract relations from web documents. In: Proceedings of KDD, pp. 712–717 (2006)
9. Zhu, J., Nie, Z., Liu, X., Zhang, B., Wen, J.: StatSnowball: a statistical approach to extracting entity relationships. In: Proceedings of WWW, pp. 101–110 (2009)

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

10. Chinchor, N., Marsh, E.: MUC-7 Information Extraction Task Definition (version 5. 1). In: Proceedings of MUC-7 (1998)
11. ACE (Automatic Content Extraction). Chinese Annotation Guidelines for Events. National Institute of Standards and Technology (2005)
12. Chen, Z., Ji, H.: Graph-based Event Coreference Resolution. In: Proceedings of ACL-IJCNLP workshop on TextGraphs-4: Graph-based Methods for Natural Language Processing (2009)
13. Ji, H., Grishman, R.: Refining Event Extraction Through Unsupervised Cross-document Inference. In: Proceedings of ACL (2008)
14. Ji, H., Grishman, R., Chen, Z., Gupta, P.: Cross-document Event Extraction, Ranking and Tracking. In: Proceedings of Recent Advances in Natural Language Processing (2009)
15. Ji, H.: Unsupervised Cross-lingual Predicate Cluster Acquisition to Improve Bilingual Event Extraction. In: Proceedings of HLT-NAACL Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics (2009)
16. Ahn, D.: The stages of event extraction. In: Proceedings of the Workshop on Annotations and Reasoning about Time and Events, pp.1–8 (2006)
17. Naughton, M. , Stokes, N., Carthy, J.: Investigating statistical techniques for sentence-level event classification. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 617–624 (2008)
18. Zhao, Y.Y., Qin, B., Che, W.X., Liu, T.: Research on Chinese Event Extraction. Journal of Chinese Information Processing 22(1), 3–8 (2008)
19. Tan, H., Zhao, T., Zheng, J.: Identification of Chinese Event and Their Argument Roles. In: Proceedings of Computer and Information Technology Workshops on IEEE 8th International Conference, pp. 14–19 (2008)
20. Xue, N.: Labeling Chinese Predicates with Semantic Roles. In: Proceedings of Computational Linguistics, pp. 225–255 (2008)
21. Surdeanu, M., Harabagiu, S.M., Williams, J., Aarseth, P.: Using Predicate-Argument Structures for Information Extraction. In: Proceedings of ACL, pp. 8–15 (2003)
22. Dorr, B.J., Zajic, D.M., Schwartz, R.M.: Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation. In: Proceedings of HLT-NAACL, W03-0501 (2003)
23. Tesnière, L.: Esquisse d'une syntaxe structurale. Klincksieck, Paris (1953)
24. Tesnière, L.: Èlm̀ent de Syntaxe Structurale. Klincksieck, Paris (1959)
25. Feng, X.: Exploration of trivalent verb in modern Chinese (2004)
26. Ningjing, L., Weiguo, Z.: A Study of Verification Principle of Valency of Chinese Verbs and Reclassification of Trivalent Verbs. In: Proceedings of 9th Chinese National Conference on Computational Linguistics (CNCCL 2007), pp. 171–177 (2007)
27. Luhn, H.P.: The Automatic Creation of Literature Abstracts. In: Proceedings of IBM Journal of Research and Development, pp. 159–165 (1958)
28. Edmundson, H.P.: New Methods in Automatic Extracting. Proceedings of J. ACM, 264–285 (1969)
29. Paice, C.D., Jones, P.A.: The Identification of Important Concepts in Highly Structured Technical Papers. In: Proceedings of SIGIR, pp. 69–78 (1993)
30. Paice, C.D.: Constructing literature abstracts by computer: Techniques and prospects. In: Proceedings of Inf. Process. Manage., pp.171–186 (1990)
31. Dali, L., Fortuna, B.: Triplet Extraction From Sentences Using SVM. In: Proceedings of SiKDD (2008)