News Event Extraction Using 5W1H Approach & Its Analysis

Smriti Sharma, Rajesh Kumar, Pawan Bhadana, Sumita Gupta

Abstract—To relieve "*News Information Overload*", a novel approach of 5W1H is easiest & best suitable. In this paper, we describe 5W1H (*who, what, whom, when, where, how*) event semantic elements extraction along with its analysis with some existing systems. Here we are also presenting a more descriptive view of 5W1H approach.

Index Terms— 5W1H, Event extraction, Semantic role labeling, NOEM

1 Introduction

SYSTEM basically relies on tools from natural language processing. The input of our system is a document with headline and text from any possible source. We implemented a gradual detection of the 5W1H with a resulting processing chain that consists of various interacting processing steps. Starting with detection of named entities, performing co-reference resolution and finally running two classifiers, to extract the WHO and WHERE, we proceed with more fine grained natural language processing on sentence level to extract the WHAT, where we use the results of the extracted WHO.

Independent of the WHO, WHERE and WHAT extraction, there are three further chains, one for each left part of the 5W1H: WHEN, WHY and HOW. The extraction mechanisms for them bear on sentence detection and pattern matching.

Recently the 5W1H concept is utilized in sentence-level understanding tasks, where it seeks to summarize the information in a natural language sentence by distilling it into the answers to the 5W questions.

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 Conferences) 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 Proteus1 is a typical MUC EE system that built for several evaluations of topics (e.g. disaster, disease outbreak) in news domain. An ACE event [2] 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 [3] 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.

- *Smriti Sharma is currently pursuing M.Tech in Computer Engineering from YMCA University of Science & Technology, Faridabad, India, E-mail: sumitagoyal@gmail.com*
- *Rajesh Kumar, Associate Professor, YMCA University Science & Technology, Faridabad, India*
- *Pawan Bhadana, Associate Professor, Maharaja Surajmal Institute of Technology, Delhi, India*
- *Sumita Gupta, Assistant Professor, Amity University, Noida, India, E-mail: sumitagoyal@gmail.com*

3 Event Semantic Extraction & collecting Event facts in news extraction

We are aiming to extract every part of the 5W1H, but not necessarily every slot can be filled, since for instance the answer to how something happened might not be given in the text.

Here 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) Collect extracted event facts automatically according to nature of existence.

3.1 Key Event Identification

We should know the answers of who?, when?, where?, what?, why?, and how? To develop a comprehensive reportage of the event.

Who

Answer to the question "Who?" is the carrier of the statement. It will tell subject of News Article .

When

Our approach to find date expressions within the text is a simple pattern matching against time and date patterns

Where

To extract the WHERE we make use of Named Entity Recognition because the place where something happens is most likely a named entity of type location which can be recognized through the NER.

What

Our WHAT describes a change of state, we decide to focus on a verb as WHAT.

Why

When it comes to the question WHY, the relational character of events becomes visible. Each event has its own internal structure, and meanwhile often relates to other events semantically, temporally, spatially, causally, or conditionally.

How

It gives detailed description of event itself.

These above elements give "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 processing chain of Who, Where, What and Why is shown in fig 1 and fig 2.

3.2 Event Semantic Element Extraction

The second step of our approach is to extract the 5W1H semantic elements for key event. We first 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. These steps also need POS-tagging, Sentence Detection , phrase chunking , Named entity recognition etc.

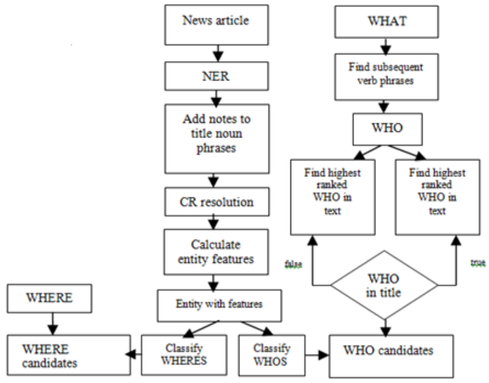
News article
NER
Add notes to title noun phrases
CR resolution
Calculate entity features
Entity with features
WHERE
WHERE candidates
Classify WHERES
Classify WHOS
WHO candidates
WHAT
Find subsequent verb phrases
WHO
Find highest ranked WHO in text
Find highest ranked WHO in text
false
WHO in title
true

Fig. 1. WHO, WHERE & WHAT processing chain

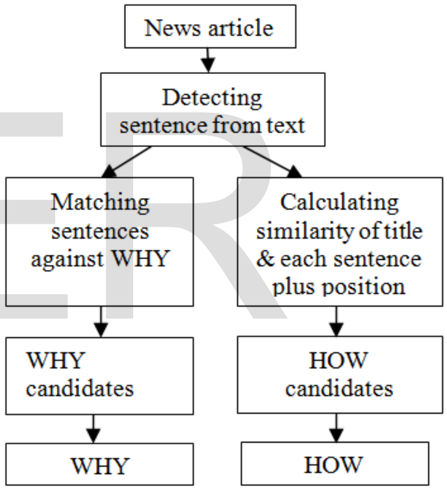
News article
Detecting sentence from text
Matching sentences against WHY
Calculating similarity of title & each sentence plus position
WHY candidates
HOW candidates
WHY
HOW

Fig. 2. WHY & WHO processing chain

3.3 Collecting Event facts , Event Extraction Process

After the text of the article is preprocessed the gradual extraction of the 5W1H starts. Here we have a general problem that the subsequent verb phrase in long sentences contains a lot of information that we cannot ignore because it is semantically relevant. For the sentence, "A court in Pakistan has sentenced a Muslim prayer leader and his son to life in jail for blasphemy." our subsequent verb phrase is "has sentenced a Muslim prayer leader and his son to life in jail for blasphemy" which indeed is correct but contains a lot more information than we need. Finally it is a non-trivial task to filter out the minimal necessary information. We decided to solve this problem by limiting the verb phrase to length. As shown in fig. 3, nodes of a parse tree helps to explain the problem of the subsequent verb phrase.

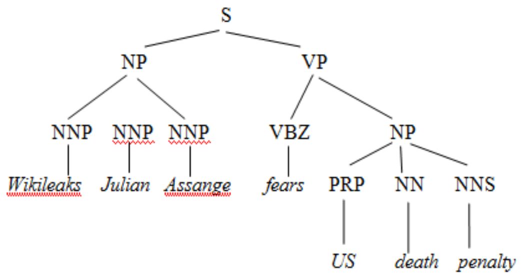

Fig. 3. Parse tree for example sentence

For this case the WHO is "Julian Assange". As one can see the list holds two candidates for the WHAT. On position 6 the verb phrase "fears US death penalty" and on position 7 in the list "fears". On position 6 the verb phrase "fears US death penalty" and on position 7 in the list "fears". Here pick out a long enough verb phrase that contains all necessary information. In the example this is "fears US death penalty". The verb "fears" here does not contain enough information to understand the message. We take the highest rated date as WHEN. addition we implemented an *NER* capable of annotating dates and time in a given text. For this we use the OpenNLP NER with only models loaded for time and date recognition. The recognized dates are added to the *rankedCandidates* for the WHEN. Next is the extraction of WHY which basically is looking for sentences that are indicating a reason for the event itself. We take the sentence with the highest confidence to fill the WHY slot. To extract the HOW from the news article we look for the sentence that contains the WHO and WHAT because this sentence most likely holds additional information about how the event happened.

4 Analysis Of systems to extract 5W1H from a news article

We explain two systems which differ in their approaches but have the common goal to extract 5W1H from a news article.

4.1 NEXUS

News cluster Event eXtraction Using language Structures (NEXUS) is and event extraction system utilized for populating violent incident knowledge bases. It automatically extracts security-related facts from online news articles. Before the NEXUS event extraction process can proceed, news articles are gathered by media monitoring software (EMM system), which delivers news clusters for each topic. Further, NEXUS selects security related events via application of key-word based heuristics. In the next step the documents in each cluster are linguistically preprocessed which encompasses the following steps: Sentence Boundary Disambiguation, Named Entity Recognition, simple chunking, labeling of action words (e.g. kill, shoot) and unnamed person groups (e.g. six civilians).

To fill the event frame describing the main event they perform an event aggregation containing the following steps:

- victims: resolving role ambiguities of recognized entities
- number of killed, wounded and kidnapped: average-like estimation
- place: geocoding via EMM system
- perpetrators: named entity recognition
- weapons: lexicon
- event type: classification over lexicon of event keywords
- date: date of the news cluster

4.2 Chinese News Fact Extractor (CNFE)

Another novel news event semantic extracting approach addressing the 5W1H based on one document was implemented in the CNFE. Since applying SRL for Event Extraction is computational intractable over large scale news corpora, they propose a "light" but effective method to extract the 5W1H. Their extraction pipeline includes three steps:

- Topic Sentences Extraction,
- Event Classification, and
- 5W Elements Extraction.

Extraction of the 5W1H is based on the extracted topic sentences. The type identification module searches the topic sentences by examining trigger list and marks each appearance of a trigger as a candidate event. From the output of event type identification and SVM Rectifier they get headline, topic sentences and a list of 5W candidates of an event, i.e. predicate, event type, named entities, time and location words. Next is to identify 5W1H semantic elements as follows:

- Who, Whom: To identify these arguments, regular expressions are used to match trigger's syntactic-semantic rules. For example, they use an expression "(.*)/n(.*)/trigger(.*?)/n(.*)/n.*" to match "NP1+V+NP2+NP3". They identify arguments from Named Entities and NPs of the trigger according to the sentence's syntactic structures. Then they determine there roles (e.g. agent, patient) and associate them with a specific template.
- What: They use the first identified Verb of their verb-driven method which is rectified by a SVM.
- Where, When: Outputs of the NER are used to identify time and location.If there are no Time/Location Named

Entities, generated chunks with tags of /nt and /ns are adopted.
• How: They combine the results to a sentence "Who did What to Whom".

They replace the Why with Whom, which enables them to rely on their topic sentences, because the answer to the question Why is scattered over the document and cannot be answered by the topic sentence. So they suppressed an important part of the original 5W1H concept, which is relevant regarding further semantic processing and setting different news into relation. On the other hand removing the why, enables
them to keep their research consistent with ACE event extraction in order to compare with other works.

REFERENCES

[1] Jakub PISKORSKI , Hristo TANEV, Martin ATKINSON,Erik VAN DER GOOT ,"Cluster-Centric Approach to News Event Extraction"

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

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

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

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

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

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

[8] 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

[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.