Business Event Extraction System Based on SSVM

Sungho Shin, Young-Min Kim, Choong-Nyoung Seon,
Seunggyun Hong, Sa-kwang Song*, Hanmin Jung

Dept. of Computer Intelligence Research, KISTI
245 Daehak-ro, Yuseong-gu, Daejeon, 305-806, South Korea
{maximus74, ymkim, wilowisp, xo, esmallj, jhm}@kisti.re.kr

Abstract. Information extraction from unstructured text data has been used essentially to provide new insights by collecting, storing, and analyzing text data in textual analysis. The research on event extraction has been recently getting more attention in information extraction area since lots of events happens and significantly affect our societies and countries. Related studies on event extraction use rules for identifying and extracting events from texts so far. However, rule based approaches have a limitation in terms of accuracy and rule construction. In this paper, we present an event extraction system which takes advantage of the machine learning method, especially SSVM. The system extracts event triggers and predefined event arguments, while existing rule based systems extract unknown event arguments. Ours provides 60.23 F1 score, which is higher than that of previous researches, in which rule based event extractions were performed. Even though rule-based and machine learning-based approaches cannot be compared against each other completely fairly, what is clear is that for the task in which event arguments are defined in advance, applying machine learning method can make better results.

Keywords: Event Extraction, Trigger, Argument, Temporal Information, SSVM, Machine Learning

1 Introduction

Recently, research on event extraction has been conducted to rapidly discover meaningful events with regard to business that are found in a massive quantity of data. In terms of meaningful information existing in text, an event might be viewed as a type of group that qualifies for information to be extracted. Data obtained from named entity recognition, relationship extraction, and event extraction can be employed as fundamental data for more detailed data analysis and intelligent services, such as natural language questioning and answering (NLQA) or predictive analytics (PA) which are emerging technologies presented in Gartner Hype Cycle [1].

Lots of researches have been conducted on named entity recognition and relation extraction, and the results have reached over 90% of human cognitive capacity [2]. In addition to the extensive works that has been performed in these areas, research on event extraction has attracted much attention recently. In particular, the prediction of

* Corresponding author

influenza spread and the assessment of the moving direction or damage status of a typhoon or tsunami, which are natural disasters, as performed by Google are the results of analyzing the extracted events. Similar to the early stages of named entity recognition and relation extraction, the research on event extraction, being still in the initial stage, has been mainly realized by rule-based methods. Known from researches on named entity recognition or relation extraction, rule-based methods have advantages of cost and time efficiency, although it considerably lacks in precision when being compared to machine learning methods. Accordingly, event extraction based on machine learning might hold greater benefits to services provided to users in actual business settings through NLQA or PA.

This study relates to event extraction as source data required to expand the knowledge database used for NLQA or PA. We aim to accurately extract a trigger and arguments of an event by using a machine learning method with applying a structural support vector machine (SSVM) algorithm. Furthermore, temporal information, another event argument, can be extracted more effectively by applying rules from a text or metadata rather than by applying the machine-learning method because of its various forms of expression. We designed and implemented a module for extracting temporal information.

2 Related Work

An event is real temporal data, whose temporal sequence is important and might be infinitely long [3]. An event can defined based on the structure of "who, when, what, where, how (5W1H)," but could consist of only some of these. In addition, occurrences among compounds in the field of biology and data transmission/reception among computer devices are also referred to as events. In particular, researchers define social events as events that occur in a relationship between people and in a society. Social events include a wide range of accidents or incidents that affect society at a small or large scale [4]. Earthquakes, typhoons, traffic congestion, dialogues, accidents, international academic conferences, etc., are considered social events. Business events such as company mergers, product launches, and company bankruptcy are included as well. A trend of event extraction research is more inclined to comprehensive event extraction research than to business event-focused research.

A variety of techniques, such as language processing, text mining, data mining, and machine learning, are widely adopted to extract events from a large quantity of text. Extraction methods are largely divided into a machine-learning method and a rule-based method. The machine-learning method sets the structure of an extraction event in advance, constructs a learning group based on the determined structure, and extracts the event. This method is advantageous for extracting events with a predetermined structure. On the other hand, although the rule-based method also defines the structure of an event in advance, the structure of the event can subsequently be changed easily, thus permitting freely reflecting feedback from the extraction result. However, there exists a limitation to the rule-based method because of a significant

number of exceptions. Accordingly, the machine-learning method might be more efficient in extracting events with a predetermined structure.

3 Event Expression

This study focuses on business event extraction reported in the literature. A business event indicates an event arising from corporate activities, and the major concepts pertaining to this type of event are as follows [5]:

- **Event mention**: an event type. Common designation of events that have an identical meaning.
- **Event trigger**: the most exact term for expressing an event mention.
- **Event argument**: an event attribute associated with an event trigger, such as an entity mention, a temporal expression, or value.
- **Event instance**: an instance comprised of event arguments associated with an event trigger that is referring to an event mention presented in an identical sentence.

According to the above definition, the event in this study includes an event trigger, event subject, event object, and time. The event subject, event object, and time correspond to an event argument. An event is expressed as shown in (1).

Event mention <event trigger, subject (arg.1), object (arg.2), time (arg.3) > **(1)**

In this study, business events are limited to 'Announce' and 'Launch'. The subjects of two events are corporation names and the objects are product names which are highly related to the companies.

4 Machine Learning based Event Extraction System

In this study, an event trigger and arguments, with the exception of temporal information, are extracted by the machine learning method. Temporal information is extracted by the rule-based method rather than by the machine learning method because such information can be expressed in various ways.

The machine learning algorithm used is the structural support vector machine (SVM) algorithm. The structural SVM algorithm is a machine learning algorithm extended from the conventional SVM algorithm. A structural SVM supports more general structural issues, for example, sequence labeling, parsing, etc., whereas a conventional SVM supports binary classification and multi-class classification, among other classification methods.

In addition, for the learning of the structural SVM algorithm, this study uses the Primal Estimated sub-GrAdient Solver for SVM (PEGASOS) algorithm among the Stochastic Gradient Descent (SGD) method as an extension of the structural SVM algorithm because the PEGASOS algorithm has shown high performance and rapid learning rate when applied to the SVM.

To extract an event by applying the extended structural SVM algorithm, training data is required. Training data can be directly constructed manually by domain experts, but because of limited time and labor, it is preferred to make it automatically. We also built it semi-automatic method [6]. The initial training data is constructed by using a simplified distant supervision method which is an automated method. For the method, seed data lists for named entity and for relation are respectively established first; Seed data is then used for key word searching to collect training sentences; sentences including the corresponding key words are extracted from the web; and a Silver standard training data is established.

Later, domain experts are hired to establish the gold standard through manual verification in order to enhance accuracy. The training data was prepared for two event types and two relation types which are just for the system evaluation. The number of training data of each type is presented in table 1.

Table 1. Types of relation & event and # of training sentences

Category	Type	Object*	# sentences
Relation	ElementOfTechnology	Technology name	1,031
	CompeteProduct	Product name	537
Event	Announce	Product name	2,448
	Launch	Product name	1,986
Total			6,002

With the machine learning, an event trigger, an event subject, and an event object are extracted; temporal information is extracted through an additional designed tool. Fig. 1 illustrates the process of extracting the event temporal information. Each module that comprises the tool is described as follows.

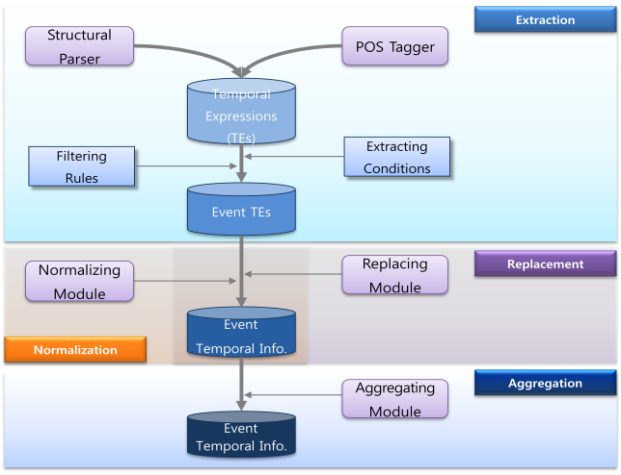

Fig. 1 Process of Capturing Temporal Information

<Extraction>

- Uses the results of parsing and POS tagging on each sentence
- Extracts temporal information from all time expression adverbs within a sentence by applying temporal information extraction conditions
- Filters or refines extracted time expression candidates according to purpose
- Excludes useless time expressions by filtering
- Deletes tokens not comprising time, such as prepositions, articles, and more, from the time information

<Replacement>

- Replaces or complements an accurate point of time by using metadata with respect to temporal information that is inferable by the metadata
- Applies temporal information replacement rules to the temporal information to be replaced

<Normalization>

- Classifies temporal information based on year-month-day combination
- Changes unstructured temporal information to a normalized form
  * Normalized form: YYYY-MM-DD
- Expresses '0000' or '00' when lacking information in the normalized form

<Aggregation>

- Assumes that an identical triple (S-P-O) has identical temporal information
- Compares each temporal information by collecting identical triples
- Aggregates temporal information lacking in the triples

5 Experiments

The evaluation of the system which is used for event extraction in this study is performed with F1 score for two business relations. The result of the evaluation is presented in table 2. Our system is well performed in information extraction job, especially relation extraction which is similar with event extraction.

Table 2. Performance of our system for relation extraction

Types	Precision	Recall	F1 score
ElementOfTechnology	87.50	89.74	88.61
CompeteProduct	90.91	58.82	71.43
Total	89.20	74.28	80.02

The test data is 10% of the entire training data (table 1), and it is randomly sampled. The result of business event extraction is shown in table 3. Even though the F1 score of each event type is different, they are all over 60.0. The F1 score of our system exceeds that of other systems [5, 7] which are based on rules (table 3). This is not absolutely true because there is a limitation that the experiment environment is not same. However, what is clear is that for the task in which trigger and arguments are fixed in advance, applying the machine learning method can make better results.

Table 3. Result of Business Event Extraction

Types	Precision	Recall	F1 score
Announce	63.18	58.53	60.77
Launch	58.76	60.45	59.68
Total	60.97	59.59	60.23

Table 4. Comparison of our system and others

Research	Trigger F1	Arg F1	Average	Method
Ji and Grishman (2008)	67.30	42.60	54.95	Rule-based
Qi et al (2013)	67.50	52.70	60.10	Rule-based
Ours	-		**60.23**	ML

6 Conclusion

This study aims to extract business events by using machine learning based information extraction system. Different with events defined in other researches, arguments of an event is predefined. Thus, our system is only capable of extracting designated arguments in advance. For extracting the temporal information, one of event arguments, we use rules. The result of system evaluation says that our system accomplishes better performance than other rule based system in terms of F1 score. This means that for the event extraction in which the arguments are defined ahead, machine learning based method makes better results.

Acknowledgement. This work was supported by the IT R&D program of MSIP/KEIT. [2014-044-024-002, Developing On-line Open Platform to Provide Local-business Strategy Analysis and User-targeting Visual Advertisement Materials for Micro-enterprise Managers]

References

1. Gartner Inc., “Hype Cycle for Emerging Technologies, 2014,” http://www.gartner.com /technology/research/hype-cycles/, 2014.
2. Bach, N., Badaskar, S., “A Review of Relation Extraction,” 2007.
3. Complex Event Processing, http://en.wikipedia.org/wiki/Complex_event_processing.
4. Sakaki, T., Okazaki, M., Matsuo, Y., “Earthquake shakes Twitter users: Real-time event detection by social sensors,” In Proceedings of the 19th International Conference on World Wide Web, pp. 851–860, 2010.
5. Li, Q., Ji, H., Huang, L., “Joint Event Extraction via Structured Prediction with Global Features,” In Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics, 2013.
6. Shin, S., Choi, Y. S., Song, S. K., Choi, S. P., Jung, H., "Construction of Test Collection for Automatically Extracting Technological Knowledge,” Journal of Korea Content Society, vol.12, no.7, 2012. (in Korean)
7. Ji, H., Grishman, R., “Refining event extraction through cross-document inference,” In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, pp. 254-262, 2008.