Event Extraction Approach from Texts

Aymen Elkhlifi[*] and Rim Faiz[**]

[*]*LALIC, Paris-Sorbonne University,*
28 rue Serpente Paris 75006
Paris, France
Aymen.Elkhlifi@paris-sorbonne.fr

[**]*LARODEC, IHEC de Carthage,*
2016 Carthage Présidence
Carthage, Tunisie
Rim.Faiz@ihec.rnu.tn

ABSTRACT. *A new challenge is added to the Natural Language Processing Community; how to analyze the new documents forms resulting from the Web 2.0? We are interested in a particular kind of information which is events. Thus, we propose a generic approach to extract and analyze events from text. We propose an event extraction algorithm with a polynomial complexity O(n^5). This algorithm is based on developed semantic map of events. We validate the first component of our approach by the development of the "EventEC" system.*

KEYWORDS: Documents and Knowledge Engineering, Information Extraction, Event Extraction, Semantic Maps, Ontology.

RÉSUMÉ. *Un nouveau défi s'ajoute à la Communauté traitement automatique du langage naturel; comment analyser les nouvelles formes de documents provenant du Web 2.0. Nous nous intéressons à un type particulier d'information qui est l'événement. Ainsi, nous proposons une approche générique d'extraction et d'analyse des événements à partir de textes. Pour cela nous avons mis en place un algorithme d'extraction d'événement qui dispose d'une complexité polynomiale O (n^5). Il se base, entre autres, sur des cartes sémantiques des développées. Nous avons validé la première composante de notre approche par le développement du système "EventEC".*

MOTS-CLÉS : *Ingénierie des Documents et des Connaissances, Extraction d'information, Extraction d'événement, Carte Sémantique, Ontologies. .*

1. Introduction

New sources of textual information, rich in events, grow significantly, such as social networks, blogs and wikis. They are added to old sources like the informative web sites, emails and forums, which shows the importance to manage these data automatically. According to the Linguistic Data Consortium, the best event extraction system allows to extract 14.44 % of the events in a textual document. This is during the last evaluation concerning the events (Automatic Content Extraction) ACE (2007). This result shows the need for re-examining the way of modeling as well as the practical strategy of event extraction. Accordingly, our research focuses on the event annotation and their analysis. First, we annotate events using an effective algorithm based on Contextual Exploration. Second, we group similar events using appropriate similarity measures of similarity. This output is very useful for many information extraction tasks like summarization, text categorization and query answering systems.

The rest of the document is organized as follows: Section (2) deals with the definition of Event and introduces the related works on event extraction methods. In section (3), we present our approach for automatic event processing, particularly the component of extracting events. The experimentation is described in section (4). Then, we evaluate the system in order to demonstrate its abilities. Finally, in section (5) we conclude our work with a few notes about the perspectives.

2. Related Works on Events Extraction

It is worth noting that the event definition varies according to the application domain: probabilities, software development, history, philosophy and linguistics. But we can be said that an event is something that happens, it can frequently be described as a change of state or a transition between two states.

ACE definition (2007) adds that an Event is a specific occurrence involving participants. However, *TimeML* specification (Pustejovsky, 2003) considers "Event" as a cover term for situations that happen or occurs. Events can be punctual or last for a period of time. *TimeML* also considered as events those predicates describing states or circumstances in which something obtains or holds true. Otherwise, Hong-Woo (2004) define Event as the binary relation between two entities for special event verbs which are predefined by biologists. Historically, the tasks of event extraction were first explored in the series of Message Understanding Conferences (*MUCs*) started from 1987. The events in *MUCs* were limited to finite topics, e.g., terrorist activities, management succession. Other recent works are hybrid (Elkhlifi et al., 2009). They use machine learning techniques to make annotation rules similar to the pattern-matching.

Several works which have followed treat the event extraction, are based on the pattern-matching rules (Mani et al., 2000), or on the machine learning approach

(Mazur et al., 2007). But the problem is the high complexity of the algorithms which are presented by these approaches. This prevents the passage on large scale.

Different systems, however, represent events in different ways. There are two approaches to represent events: On the one hand, there is the *TimeML* model, in which an event is a word that points to a node in a network of temporal relations. On the other hand, there is the *ACE* model, in which an event is a complex structure, relating arguments that are themselves complex structures, but with only ancillary temporal information. In our study, we are interested rather in the annotation of the events in the form of metadata on the document; we propose our ontology of events and our method to extract them.

3. Our approach

Our model of event extraction is a component in a broader approach that we propose for processing and analyzing events. This approach is composed of the following parts:

– A first component of extracting and clustering events: starts with the segmentation of text. Then, the annotation of the events using the Contextual Exploration technique.

– A second component of analyzing event clusters by a Categorical Applicative Grammar '*CAG'*. In a first stage, we generate the phenotype configuration. Then, we determine the normal form of event (the operator/operand structure). This structure is the semantic functional form of event. We propose to develop a '*Heuristic CAG'*, a new version of *CAG*, where we suppose some constraints on the type initially affected.

– A third component of the exploitation of the events by storing the normal form in a relational database. For that, we determine a procedure which describes the transformation of normal form into database schema. Information which we want to fill in the database is mainly: Situations (state, processes, event, resulting state, etc.), Agents and Circumstances (spatial and temporal).

In this article, we present the first component of the general approach described below. We will present the first part and its experimentation independently of the other parts. We initially segment the text into different units. Then, we propose an algorithm which annotates the events efficiently. The efficiency is represented by a minimal complexity compared to the other algorithm described in the literature.

3.1. Ev*ent Extraction: Segmentation*

The segmentation is the determination of the unit's borders (unit as sentences, paragraphs etc.). It is a hardly-realizable task. Given that a point followed by a capital letter is not enough to detect the end or the beginning of a segment, it is necessary to take into account all typographical markers. Moreover, other linguistic bases are engaged like the syntactic structure of a sentence and the significance of

each typographical marker in a well defined context. The existing tools segment the structured texts into paragraphs. But, the segmentation of texts in smaller units (sentences) remains a difficult task currently. There exist some works related to the monolingual segmentation, in French language, English and German. We developed our own segmentor while basing ourselves on punctuation marks. (Elkhlifi et al., 2007).

3.2. *Event Extraction: Annotation using Contextual Exploration*

Contextual exploration 'EC' is an effective technique (Desclés et al., 1997). It takes into account the context to commit semantic indeterminations or to make decisions in the construction of meaning. It lies within the scope of rule-based methods in Artificial Intelligence. It consists in applying rules in a context which is determined by indices (hierarchized indices: first, indicators and secondly, complementary clues). EC has the advantage of being independent of the application field, because the rules describing the linguistic phenomena are independent of a particular area. In addition, it doesn't need a morpho-syntactic analysis. This factor reduces considerably the execution time when we implement the method. Event extraction can be seen as a discursive point of view in information extraction and it is indicated by linguistic markers of surfaces (verbs, nouns and adjectives). Some indicators are polysemous, thus they need a complementary clues to clarify the indetermination.

We define an event as a fact which occurs at a given time. It can be punctual or continuous. An event is characterized by a transition between states. We present the event in general as aspectual information which can be identified by linguistic markers: verbal expressions (such as the occurrence verbs), noun expressions (*the death of X*) or some adjectival expressions. An event is announced by one or more reporters. The event occurs at a well defined time in a specific place. But these two attributes are optional. We can announce an event, without giving the place or time. An event can be carried out or not carried out. That leads us to define specific clues of times. We observe the succession of several events in a text. They are inter-related, and we are interested mainly in the relation of causality between them.

We defined the semantic map for a particular field which is the natural disasters: a disaster has several types and is caused by climatic changes or other factors. It causes human and non human damages (see figure 2). Our choice is explained by the richness of this field in event and their diversities. This semantic map can be seen like a linguistic ontology which is going to be re-used by other ontology. To annotate events, we propose the following algorithm:

1: Let E a set of rules defined for the semantic map of the events (Ontology of Event)

2: $E = \bigcup_{n}^{1} R_i, R_i \{I_i, C_{pi}, C_{ni}, C'_{pi}, C'_{ni}\}$

3: With R_i represent an event annotation rule having I_i as an indicator,
C_{pi} ,C_{ni} respectively their clues **right-hand** side ***positive*** and ***negative***.
C'_{pi} ,C'_{ni} respectively their clues **left-hand** side ***positive*** and ***negative***.

4: Let I a set of indicators relating to E, $I = \bigcup_{n}^{1} I_i$

5: Let D a document, and S the set of segments that we can form of D.

$S = \{S_i; \bigcup_{n}^{1} S_i = D\}$ et S_i = {Propositi on, sentence, paragraph, etc.}

6: $S_i = \{T_i; \bigcup_{n}^{1} T_i = S_i\}$ the set of the terms forming S_i

7: For each S_i
8: If $(\exists T_i \in I)$ Then
9: $F = \{R_i; T_i = I_i\}$ a set of rules having T_i as an Indicator
10: Let C_L, C_R respectively **Right-hand Context** and **Left-hand Context** of S_i
L = Left or right part starting from T_i
11: For each R_i
12: CluesPositive = False; CluesNegative = False;
13: If $\neg((\exists T_J \in C_L) \wedge (T_J \in C'_{pi})) OR (\exists L_i \in C_{ij})$ Then
14: CluesNegative = True;
15: ElseIF $((\exists T_J \in C_R) \wedge (T_J \in C_{pi})) AND ((\exists T_J \in C_R) \wedge (T_J \in C_{ni}))$ Then
16: CluesPositive = True;
17: End if
18: End if
19: If (CluesPositive × CluesNegative = true) Then
20: ***Annotate (S_i, R_i)***
21: End if
22: End for
23: End if
24: End for

Figure 1. Event Annotation Algorithm

The algorithm of event annotation takes as input a semantic map and a set of rules, to annotate the event efficiently. If we consider that the basic unit is equal to the comparison of two patterns, then the complexity of our algorithm is $O(n^5)$, with n the number of segments which can be formed from a Document. Whereas the complexity of the algorithm proposed by (Bittar, 2008) for French language is $O(n^{11})$, it uses the full analysis of sentence.

Taking for example this title "*Avalanche au Kirghizistan: 5 morts - Avalanche in Kirghizistan: 5 dead*". We notice that the authors tend to express the events in a

short way on the titles' level. This is why; we define specific rules for the titles. In the example above, the title contains two events connected between them: **Event 1**: *"Avalanche"*, **Event 2**: *"5 morts- 5 dead"*, **Relation** between event 1 and 2: causality relation expressed by two points.

For the Avalanche class on the title level, it is enough to find an occurrence T_i belonging to the avalanche indicator to annotate the segment as an "Avalanche event". The nominal indicator of this class is the word "*Avalanche*" and these synonyms like "*Masse de neiges - snow mass*" and "*bloc de neige snow block*" etc. We expressed this by a regular expression.

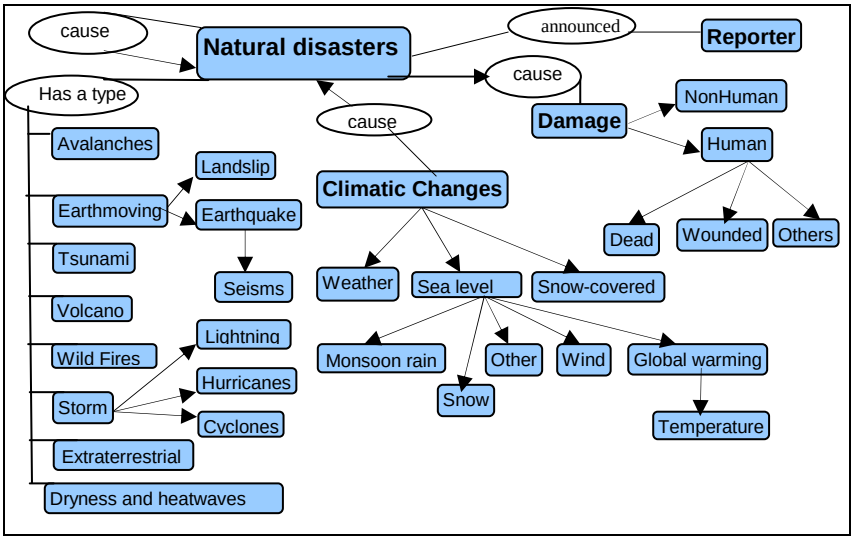

Figure 2. Map of Natural Disaster

Beyond the title, the existence of an avalanche indicator does not imply an event. We must seek indices with the periphery indicator. It becomes an event if we find a verb of occurrence, such as for example this sentence: **Event 3**: "*…une avalanche qui **s'est abattue** sur la…- …an avalanche which **stroke** the area…":* In addition, if the avalanche is dated then it is also an event for example: "*L'avalanche de **jeudi** the **Thursday** avalanche*", or "*L'orage de **l'année 2000** - **The 2000** storm*".

Therefore the rule which expresses the example above is mentioned below:

If $\exists$ an occurrence $T_i \in I_i = I_{Avalanche}$

If $\exists$ an occurrence $Y \in C_{production}$

OR If $\exists$ an occurrence $Z \in C_{Times}$

it

ıı Then Annotate the segment container T_i as an Avalanche Event nominal (d).

a) *"Ils sont tous mort* - They are all died*"*.
b) *"Il vient de décéder* - He has just died".

c) *"Un orage a tué deux personnes* - A storm left 2 people dead*"*
d) *"La mort inattendu de 17 personnes* - The unexpected death of 17 people*"*

We express them respectively by the following expressions

a) (est| sont| était| étaient| fut| furent) (mort(s)?| décédé(s)?)
b) (vient| viennent) (de|d') (mourir| décéder| expirer| cesser| périr| emporter| succomber| trépasser)
c) (tu| cess) (e| es| ons| ez| ent| ais| ait| ions| iez| aient| ai| as| a| âmes| âtes| èrent)
d) Lists of nouns indicate "mort" with an article.

We defined rules for each group of indicators. If the verb is in the past or past simple, it expresses an event in French language. If not, we must seek other indices which confirm the event like the dates and the places. Some verbs do not imply an event only with the 3rd and them pronoun (like to die). That's why we filter all erroneous forms in the expression of indicator.

The first semantic natural disasters map was used to understand event in a specific field, our objective is to extract them in general. We defined for that the generic map. An event can be social, individual or natural; the social event can be Economic, Cultural, Conflicts, Legal or others. For each class we have to determine its sub-classes as follows:

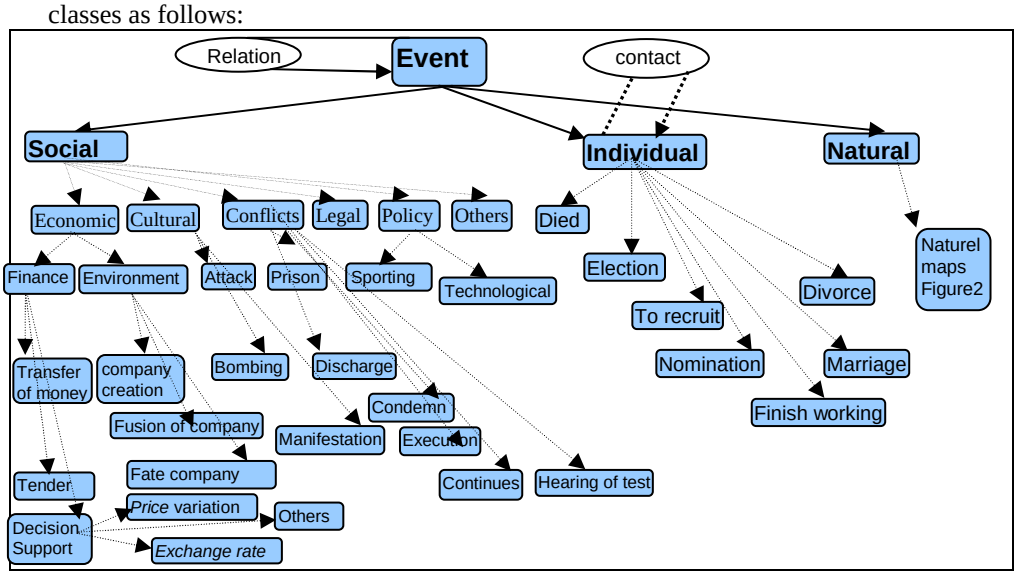

Figure 3. Semantic map of Events

For each concept of the map, we defined the set of rules which covers all the possible linguistic forms of event. We have developed about 200 rules. We start from a textual example to generalize all linguistic manifestations. This method makes it possible to define incrementally a solid base of rules.

4. Experimentation and Results

To validate our model, we develop the *EventEC* system in Java using Eclipse environment. *EventEC* includes the module of segmentation and Event Extraction. We prepared a corpus containing 753 articles from many sources: Blog: 117 articles, Wiki: 185 articles, News articles: 256 articles, Social Web: 96 articles; Email: 102 articles. The average length of a sentence is of 11.54 words, with an average of 6.1 events per document, for a total of approximately 252054 words, 21837 sentences and 4594 events. This corpus was annotated by tow experts. For each segment of the article, they indicate whether it represents an event or not. If yes, they affect a class from the semantic map to the segment.

After removing the images and the legends of the articles, we segment them into sentences and we apply our algorithm of event annotation.

To evaluate this algorithm, we employ the following definition of the precision and the recall.

- *a*: Recognized by the system and the annotator
- *b*: Not recognized by the system but annotated by the annotator.
- *c*: Recognized by the system but by not annotated by the annotator

The Precision and the Recall is calculated as:

$$P = \frac{a}{a+c} \qquad R = \frac{a}{a+b} \qquad F1 = \frac{2 \times P \times R}{(P+R)}$$

We obtained the following value for precision and Recall: P = 82.314%, R = 89.831% and F-sore = 85.91.

4. Conclusion

In this paper we proposed a model of event extraction which is based on Contextual Exploration. We have proposed a polynomial algorithm to annotate events. We developed a semantic map of events, and a set of rules which are associated to each concept of the map. Also, we developed the *EventEC* system composed of two modules in order to evaluate the model. This work comes within the framework of the extraction and the exploitation of the events. Actually, it constitutes a considerable target in many application domains like national security, economy and biology. In short term, one of the first future works which we propose is to analyze the obtained clusters of events by GAC.

In long term, we look forward to fuse the events. In effect, we have the idea of adopting, to the case of the events, the MCT model for the fusion of information in general.

References

ACE., English Annotation Guidelines for Events. 2007. ACE report by the Linguistic Data Consortium. http://www.ldc.upenn.edu/Projects/ACE/

Bittar A., « Annotation des informations temporelles dans des textes en français », *In RECITAL 2008*, Avignon, juin 2008.

Boguraev B., Ando R-K., « TimeBank-Driven TimeML Analysis », *In Annotating, Extracting and Reasoning about Time and Events*, 2005.

Demeure I., Farhat J., « Systèmes de processus légers : concepts et exemples », *Technique et Science Informatiques*, vol. 13, n° 6, 1994, p. 765-795.

Desclés J.P., Jouis C., Reppert D., « Exploration contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte », *In D. Herin-Aime and alii (eds) Knowledge modeling and expertise transfer*, p.371-400. 1991

Elkhlifi, A., Faïz, Rim. « *Machine Learning Approach for the Automatic Annotation of Events* ». Proceedings of the 20th International FLAIRS 2007, AAAI Press, pp 362-367.

Elkhlifi, A., Faïz, Rim. « *French-Written Event Extraction based on Contextual Exploration*». Proceedings of the FLAIRS 2010, Florida, USA, May 2010, AAAI Press. pp 189 – 198.

Hacioglu K., Chen Y., Douglais B., « Automatic time expression labeling for english and chinese text », *In CICLing*, 2005, p. 548–559.

Hong-Woo C., Ohta T., Kim J.D., Tsujii J., « Building Patterns for Biomedical Event Extraction », *In the 15th Internat conference on Genome Informatics GIW 163-164*. 2004.

Mani I., Wilson G., « Processing of News », *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000)*, p. 69–76, 2000.

Mazur P., Dale R., « The DANTE temporal expression tagger », *In Proceedings of the 3rd Language & Technology Conference (LTC)*, Poznan, Poland. 2007.

Mourad G., « La segmentation de textes par Exploration Contextuelle automatique, présentation du module SegATex », *In Inscription Spatiale du Langage : structure et processus ISLsp*, Toulouse. 2002.

Pustejovsky J., Castano J., Ingria R., Sauri R., Gaizauskas R., Setzer A., Katz G., Radev G., « TimeML: Robust specification of event and temporal expressions in text », *In AAAI Spring Symposium on New Directions in Question-Answering*, p 28–34, Stanford, 2003