French-Written Event Extraction Based on Contextual Exploration

Aymen Elkhlifi[1] and Rim Faiz[2]
[1]LALIC, Paris Sorbonne University, 28 rue Serpente Paris 75006, France
Aymen.Elkhlifi@etudiants.univ paris4.fr
[2]LARODEC, IHEC de Carthage, 2016 Carthage Présidence, Tunisie
Rim.Faiz@ihec.rnu.tn

Abstract

Event extraction is a significant task in information extraction. This importance increases more and more with the explosion of textual data available on the Web, the appearance of Web 2.0 and the tendency towards the Semantic Web.

Thus, we propose a generic approach to extract events from text and to analyze them. We propose an event extraction algorithm with a polynomial complexity $O(n^5)$, and a new similarity measurement between events. We use this measurement to gather similar events.

We also present a semantic map of events, and we validate the first component of our approach by the development of the "*EventEC*" system.

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 evolution concerning the events (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 extraction and their analysis. First, we extract events using an effective algorithm based on Contextual Exploration. Second, we group similar events using our measure of similarity. This output is very useful for many information extraction tasks like summarization, information retrieval and text categorization.

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 and grouping the similar 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.

Related Works on Events

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. ACE definition adds that an Event is a specific occurrence involving participants (ACE 2007). Whereas TimeML specification consider Event as a cover term for situations that happen or occur (Pustejovsky et al. 2003). Events can be punctual or last for a period of time. TimeML also consider as events those predicates describing states or circumstances in which something obtains or holds true.

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.

Several works which have followed touch the event extraction, are based on the pattern-matching rules (Mani and Wilson 2000), or on the machine learning approach (Boguraev and Ando 2005). But the problem is the high complexity of the algorithms which is presented by these approaches. This prevents the passage on large scale.

Other recent works are hybrid (Elkhlifi and Faiz 2009). They use machine learning techniques to make annotation rules similar to the pattern-matching (Elkhlifi and Faiz 2007).

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.

TimeML is a language of annotation, based on three concepts: Times, Events and Relations. It recommends detecting the temporal expressions according to the TimeX3 standard; then to classify the events in one of the seven suggested classes, and to determine the relations between events. These relations can be temporal, subordination or aspectual.

The Automatic Content Extraction (ACE) program starting from 1999 extended the event extraction task to 8 event types (33 subtypes) from much wider sources. The ACE program defined the following terminology for event extraction task:

- Trigger: the word that most clearly expresses an event's occurrence
- Argument: an entity mention, a time expression or value that plays a certain role in the event instance
- Event mention: a phrase or sentence with a distinguished trigger and arguments

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.

Our Approach

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

- A first component of extracting and clustering events: start with the segmentation of text. Then, the annotation of the events using the Contextual Exploration technique. After that, we gather the similar events into clusters.
- 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 resulting, etc), Agents and Circumstances (spatial and temporal)

  The last part of this component consists in enriching stored information by new knowledge relating to the normal form structure. This knowledge comes from new texts, in order to build cartography of this structure. The obtained cartography (a semantic map) can be viewed as a linguistic ontology of the events. This ontology can be used by a Query-Answer system to reply to questions about events like: Who? How? When? Where?

We will be able to answer questions like:

- Who are the actors implied in such event?
- Which are the consequences of a given event?

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. Finally, we group similar events.

Event Extraction: Segmentation

The segmentation is the determination of the unit's borders (unit as proposition, 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, English, and German language. Other more recent works considered the multilingual aspect, like the work of (Mourad 2002) which proposed an approach that consists in defining a textual segment starting from a systematic study of the punctuation marks. We developed our own segmentor while basing on punctuation marks. Due to the great number of the linguistic rules to program, we have to integrate in our knowledge base all the rules developed in the Segatex system.

Event Extraction: Annotation using Contextual Exploration

Contextual exploration 'EC' is an effective technique (Declés 91; 97; 2006). 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 was validated by (Djioua et al 2006) and (Alrahabi and Deslés 2008).

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.

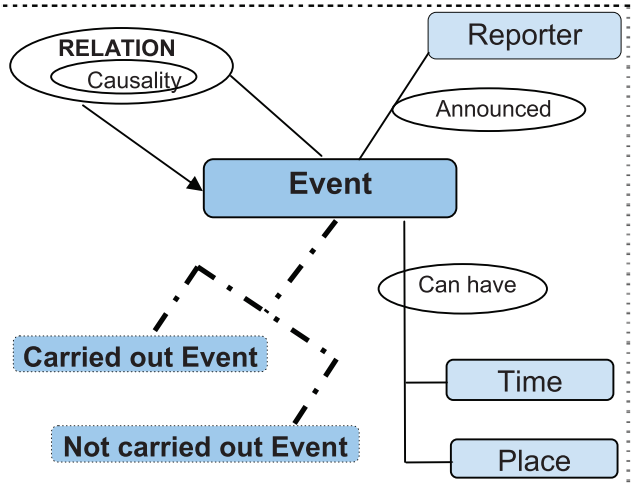

Figure 1. *presentation*

As the figure 1 shows above, 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 event, we propose the following algorithm:

Let E a set of rules defined for the semantic map of the events.

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

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

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

$S = \{S_i ; \bigcup_{n}^{1} S_i = D\}$

S_i= *{Proposition, sentence, paragraph, etc.}*

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

For each S_i

If $(\exists T_i \in I)$ Then

$F = \{R_i ; T_i = I_i\}$ a set of rules having T_i as Indicator

L = Left or right part starting from T_i

For each R_i

CluesPositive = True, CluesNegative = False,

If $(\exists T_i \in C_{ij})$ then CulesPositive =True

If $\neg(\exists T_i \in C'_{ij})$ then CulesNegative =True

If (CulesPositive × CulesNegative = true)Then

Annotate (S_i, R_i)

End if

End for

End if

End for

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.

Taking an example:

Avalanche au Kirghizistan : 5 morts
Avalanche in Kirghizistan : 5 dead
Cinq personnes ont trouvé la mort dans une avalanche qui s'est abattue sur la région d'Issyk Koul au Kirghizistan, a annoncé l'antenne régionale du ministère kirghiz des Situations d'urgence.
Five people died in an avalanche which stroke the area of Issyk Koul in Kirghizistan, said the regional antenna of the kirghiz ministry of Emergencies.
Les secouristes, aidés par les villageois, ont pu retrouver trois corps ensevelis sous la neige.
Les recherches se poursuivent, a indiqué la source…
The first aid workers, helped by the villagers, found three bodies buried under snow. Research is underway, indicated the source...

***Figure 3.**: Extract of the article "Avalanche au Kirghizistan ". Le monde 2009*

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.

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 the first sentence in the example:

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 $\exists$ *n occurrence* $T_i \in I_i = I_{Avalanche}$
If $\exists$ *n occurrence* $Y \in C_{production}$
If $\exists$ *n occurrence* $Z \in C_{Times}$
Then Annotate the segment container T_i *as an Avalanche event*

With regard to ***Event 4***: *"Cinq personnes ont trouvé* ***la mort"***, Five people **died**, It has *"la mort"* as indicator and *"Cinq personnes" as a* clues.

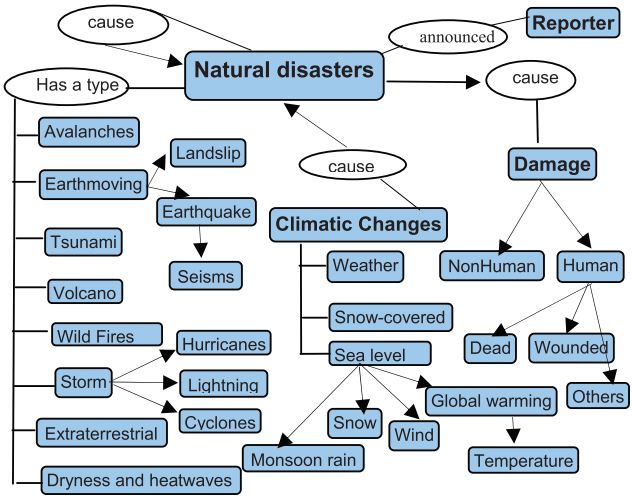

***Figure 2.**: Semantic map of Natural Disaster*

Generally the indicator of class of the class "*mort*" died can be adjectival (a), or verbal (b,c) or nominal (d). Examples:

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) (del'd') (mourir| décéder| expirer| cesser| périr| emporter| succomber| trépasser)
c) (tu| cess) (el 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.

- ***Event 6***: "*a annoncé l'antenne régionale du ministère kirghiz…*" said the regional antenna of the kirghiz ministry...
- ***Event 7*** : *"a indiqué la source"* indicated the source

Events 6 and 7 are of type Reporter. We divided this class into two sub-classes: Committed Reporter: indicated by verb like *"estime, crorit etc"* (believe, estimate). And non-committed Reporter where he takes a distance from the enunciation expressed by *"Selon, indiqué par, etc"* (According to, indicated).

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:

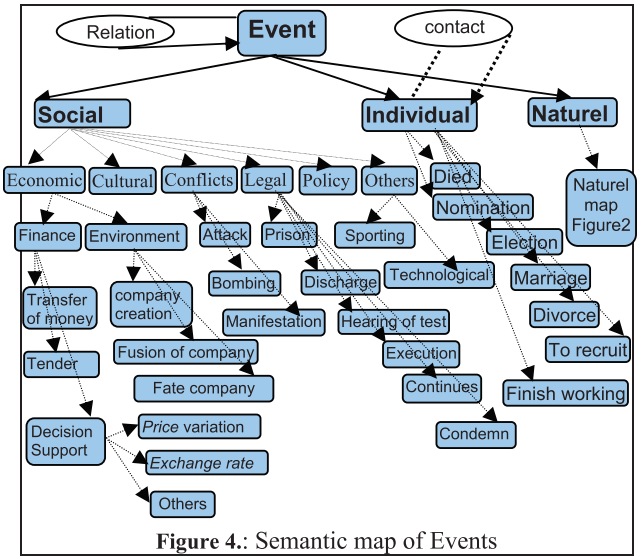

Figure 4.: Semantic map of Events

For each concept of the map, we defined the set of rules which covers all the possible linguistic form 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.

From this semantic map and the defined sets of rules, we apply our algorithm to extract the events; we describe thereafter how to group the similar events into clusters.

Clustering Event

In this stage, we gather the sentences referring to the same or similar events by the application of the algorithm 'Hierarchical Agglomerative Clustering *HAC*'(Liu et al. 2005). This algorithm initially assigns each object with a cluster, then collects on several occasions the clusters until one of the stop criteria is satisfied.

Our contribution consists in putting forward a new similarity measurement between the events. Given the importance of similarity measurements in clustering, we noted that there are several of such measurements between documents or sentences: Manhattan or Minkowski distance, Salton's cosinus and Khi-Deux distance.

We put forward a new similarity measurement between events inspired from tf-idf 'weight term frequency inverse document frequency'. This measurement also takes account of the clusters position in the document.

In order to gather sentences expressing the same or similar event by two different lexicons, we use a synonyms database for the replacement of the instances by their classes. For example, let us have the two following event-sentences, initially considered as two clusters C_1 and C_2.

- **C_1**: *"À Baqouba, deux incidents de tir séparés ont laissé six morts et ont été blessés ce dimanche après midi."*

  In Baquba, two separate shooting incidents left six dead and 15 wounded Sunday afternoon.

- **C_2**: *"Deux bombardements de voiture dans le nord de la ville de Kirkuk ont tué 10 et ont blessé 32 personnes, et une explosion dans la ville de Bassora en a tué cinq et en a blessé 15."*

  Two car bombings in the northern city of Kirkuk killed 10 and wounded 32, and a blast in the city of Basra killed five and injured 15.

We notice that the words (bombardments and bombings), (wounded and injured) imply the same meanings. Hence, there is a need to replace these words by their classes from the synonyms database in order to increase the similarity between both clusters. In general, the similarity between two classes expressing the same or similar event by means of two different lexes. We define SIM between two clusters C_1 and C_2, then, as follows:

$$SIM\ (C_1, C_2) = \frac{\sum_{j=1}^{t} Ct_{1j} Ct_{2j}}{\sqrt{\sum_{j=1}^{t} Ct_{1j}^2 + \sum_{j=1}^{t} Ct_{2j}^2}}$$

With Ct_{ij} as the weight of each term in a cluster after the replacement of instances by their classes from synonyms database. It is calculated as follows:

$$Ct_{ij} = tf(t_i, c) \times \log\left(N / df(t_i)\right)$$

- *tf(t_i , c)* the frequency of the term t_i in a cluster *c*.
- *N* the number of clusters.
- *df (t_i)* the number of clusters containing the term t_i.

We express the position of a cluster in an article as follows:

$P(Ct_i) = \frac{Order(Ct_i)}{NbCluster}$ with *Order (Ct_i)* is the position of the cluster in the document, and *NbCluster* is the total number of clusters.

Based on what has been said in so far, we propose the new similarity measurement *FSIM* which combines the similarity between clusters and the distance between them:

$$FSIM(C_1, C_2) = \alpha \times SIM(Ct_1, Ct_2) + (1-\alpha) \times D(Ct_1, Ct_2)$$

With *D (Ct_1, Ct_2)* the distance between both clusters in the article and $\alpha \in$), 1] fixed during the experimentation. Therefore, for *N* clusters, we have *n×(n-1)/2* possible combinations. It is important to group the sentences indicating the same or similar events since they will be gathered even if they use various words

Experimentation and Results

To validate our model, we develop the *EventEC* system with Java language under Eclipse environment. EventEC includes these two following modules:

- Module 1: The segmentation and Event Extraction
- Module 2: Event Clustering.

We prepared a corpus containing 753 articles from many sources (blog: 117 articles), (wiki: 185 articles), (news

articles: 256 articles), (social web: 96 articles) and (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. We obtained the following value for precision and Recall: P = 83%, R=81%

Later than the extraction of event, we use a synonyms database. Besides, we annotate the events and we group them according to their similarities. We develop several interfaces to ensure the management of corpus.

To evaluate the method of clustering, we employ the definition of the precision and the recall proposed by (Hess and Kushmerick 2003). We assign each pair of sentences in one of the four following categories:

- *a*: Grouped together (and annotated like referring to the same event).
- *b*: Not grouped together (but annotated as referring to the same event).
- *c*: Grouped inaccurately together.
- *d*: Correctly not grouped together.

The Precision and the Recall is calculated as:

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

We obtained an improvement of Recall (R) and Precision (P) and the function F1

R = 85%, P=87% and F1=73.33%.

This improvement is made to the semantic measurement of similarity which we developed. Indeed it detects the similarity between the sentences even if it contains different terms.

Conclusion and future Work

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, and a new measurement of similarity to gather those which are similar into clusters. We also 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 the extraction and the processing 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.

References

ACE 2007. *Automatic Content Extraction, English Annotation Guidelines for Events*. Report of the Linguistic Data Consortium (eds.). 2007.

Alrahabi M., Desclés J.P. 2008. *Automatic annotation of direct reported speech in Arabic and French, according to semantic map of enunciative modalities*. In 6th Inter Conf on Natural Language Processing, Gothenburg, Sweden, pp 41 51

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

Desclés, J P., Jouis, C., Oh, H., Reppert, D. 1991. *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, 371 400.

Desclés, J P. 1997. *Systèmes d'exploration contextuelle*. In C. Guimier (ed.) Cotexte et calcul du sens, Presses Universitaires de Caen, 215 232.

Desclés J P. 2006. Contextual Exploration Processing for Discourse Automatic Annotations of Texts. In FLAIRS 2006, invited speaker, Melbourne, Florida, pp 281 284

Djioua B., Flores, J.G., Blais, A., Desclés, J.P., Guibert, G.,Jackiewicz, A., Le Priol, F., Nait Baha, L., Sauzay, B. 2006. *EXCOM: an automatic annotation engine for semantic information*. In FLAIRS 2006, Melbourne, Florida, 11 13 mai, Melbourne, Florida, pp 285 290

Elkhlifi, A., Faiz, R. 2009. Automatic Annotation Approach of Events. International Journal of Computing and Information Sciences (IJCIS), Vol. 7, No. 1.pp 40 50

Elkhlifi, A., Faiz, R. 2007. Machine Learning Approach for the Automatic Annotation of Events. In the Proceedings of the 20th International FLAIRS 2007. D. Wilson and G. Sutcliffe (Editors), AAAI Press, California, , 362 367

Faiz, R. and Elkhlifi, A. 2009. Annotation sémantique des événements, in "Annotations automatiques et recherche d'informations. Hermes Traite IC2 Serie Cognition et Traitement de l'information.

Liu, H., Yu, L. 2005. *Toward Integrating Feature Selection Algorithms for Classification and Clustering*. In IEEE Transaction on Knowledge and Data Engineering, 2005. vol. 17, n. 3.

Mani, I., and Wilson G. 2000. *Processing of News*. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics pp. 69 76.

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

Pustejovsky, J., Casta˜no, J., Ingria, R.,. Saur´ı, R,. Gaizauskas, R., Setzer, A., Katz, G. and Radev, D. 2003. TimeML: *Robust specification of event and temporal expressions in text*. In AAAI Spring Symposium on New Directions in Question Answering, pp. 28 34.