

EXEV : Extracting events from news reports

Rim Faiz

AMID Laboratory – Institut des Hautes Etudes Commerciales* – 2016 Carthage_Présidence –
Tunisie – Rim.Faiz@ihec.rnu.tn

Abstract

The Press is one of the most used documentary sources. It distributes various information and presents a huge amount of data conveying a very particular knowledge type which is the event.

Our work, which fits in this frame, consists in processing, analyzing and automatically extracting textual information from electronic documents of Press based on the event as its fundamental notion.

Our approach uses a fine specific description of event and is based on a methodology of event classes and on a contextual exploration method. This method is validated by the EXEV system that we have achieved. This system analyses the electronic texts of Press in order to extract factual information with a view to providing a survey regrouping the main contemporary events.

Keywords : Filtering, Automatic Information Extraction, event classes method, Natural Language Processing.

1. Introduction

Because of the important role of information in solving many developments, its extraction and its synthesis, i.e. its processing has become of big concern. This particular interest aims to emphasize information instead of having to look tediously of it in the whole document (Faiz, 1999). This kind of processing has been done on several types of information such as the temporal information (Faiz, 1998), the causal (Garcia and al., 1999) and the defining one (Cartier, 1998).

The Press in one of the most used documentary sources. It distributes particular and diverse information and enormous amount of information conveying a very particular information type called the event.

Our work which fits in this frame, consists in processing, analyzing, filtering and extracting automatically some textual information from electronic documents of Press taking the event as its fundamental notion. This research requires a considerable effort on behalf of the cognition. On the other hand it becomes almost impossible to proceed by hand because of the considerable number of Press documents.

Our objective, in this research is to skip this excessive useless information by means of filtering and extracting in order to emphasize the event information type. As a result, the reader in this research will be urged to look for relevant information, and the journalist will be helped in developing articles surveys representing the main events.

The representation of information signaling the presence of "an event" is such an important task in Artificial Intelligence as well as in natural language processing. Indeed, just as the

* This study is realized in collaboration with the CAMS Laboratory, University of Paris-Sorbonne, France.

reasoning from information presented in a text, the understanding process must also allow the re-building of the structure of event information.

Our approach uses a fine specific syntactic description of events and is at the same time based on the methodology of event classes and on the contextual exploration method (Desclés, 1997). This method is validated by the EXEV system (event extraction) that we have achieved. The latter analysis electronic texts from Press articles in order to extract Event information, its further aim will be to provide a survey which groups together the main contemporary events.

In the following article, we will describe the EXEV system that we achieved. In order to validate our survey, we will also list the different progressive steps we followed to carry it out.

2. Presentation of the EXEV System

The system aims at automatic filtering of significant sentences bearing information with factual knowledge from Press articles as well as identifying the agent, the location, and the temporal setting of those events.

Our system (Cf. Figure 1) which relies on the JAVA language is divided in five modules:

- 1) A lexical analysis module allowing the chunking of a text into sentences and into words.
- 2) A morphological analysis module identifying words while triggering functions that deal with morphological inflexions and generate a morpho-syntactic code for each word.
- 3) A syntactic analysis module that re-establishes the order of the morpho-syntactic codes generated by the morphological sensor; and this, with the aim of building some morpho-syntactic structures.
- 4) An extraction module which allows us to pick out markers in order to identify the distinctive sentences which represent events.
- 5) A module for interpretation of the sentences which are extracted to identify "Who did what?" "to whom?" and "where?".

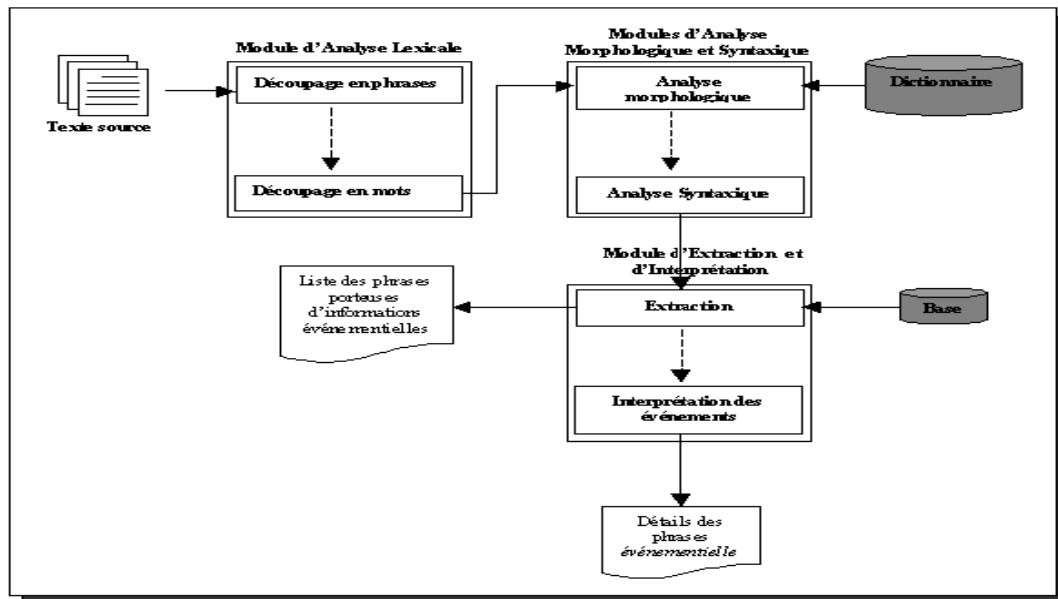


Figure 1: Architecture of the EXEV system

3. The different phases of the realisation of the EXEV System

We will briefly introduce the different phases of the realisation of our system :

3.1. Lexical analysis

It is an essential module for each text analysis whatever its type may be. Its fundamental task is segmentation. Our system splits the document into sentences and then into words by means of very precise definer detection. Example : “,” “?” “!”. We also planned to have functions which would deal with every case of ambiguity or nuance such as acronyms and abbreviations.

The role of the lexical marker is therefor to provide the “raw” basis to have access to the dictionary of the morphological analyser and this will help the recognition of words.

3.2. Morphological analysis

The purpose of this analysis is to recognise the lexical unit provided by the lexical analyser and to locate the linguistic data stored in the dictionary such as genre, syntactic category, etc.)

This recognition requires, in the first place, the calculation of a possible valid inflexion starting from the raw basis which provides two variables : basis and inflexion respecting the following condition : **word = basis + inflexion**. The basis variable represents the key for dictionary research. This, will be loaded in a chunking table to save time for reseach. Once we identify the basis, we will have information about syntactic category, root and inflectional model.

The inflexion variable, on the other hand, allows us to calculate other variable which are the following : verd tense, verb form, type and number for names and adjectives. The calculation will be possible if we refer to the inflectional model that has already been spotted thanks to the root variable.

When we finish the morphological analysis, we will have a syntactic category for each lexical unit such as genre, number, root and verb tense. To extract factual information, we only need the root, the verb tense and the syntactic category, though we could extend the calculation process to other variables such as genre or number.

3.3. Syntactic analysis

The syntactic analysis of natural language cannot directly achieved by means of the linguistic area. It needs regrouping rules. These rules imply that we first categorise the text form we intend to analyse; and this is the purpose of a morphological analysis.

On the other hand, it is worth mentionning that the syntactic analysis cannot be efficient unless it uses prediction as much as possible. This can occur because of the morpho-syntactic structure. Those structures are sometimes signalled by their form. Very often, they are spotted thanks to their precise morphological features. Their location is possible through the morphological analysis of the corresponding shapes (or forms).

The maximal analysis frame that we can logically adopt is the sentence. These seems logical since our task is to extract sentences. Therefore, the sentence is the maximal item for analysis.

Starting from the morpho-syntactic structures, which are the result of the morphological analysis, the syntactic analyser will order and build sentences with structures that will allow us to apply the extraction process.

3.4. Automatic extraction of factual sentences from Press articles

3.4.1. Sentences representing the corpus

As mentioned above, we choose press articles as our corpus. Those articles supply the reader with information which he will have to analyse. In this section, we give examples of sentences automatically extracted by the EXEV system from an electronic document of press : *Le monde 12 February 1999* :

- 1) *The murderous avalanche which hit the valley of Chamonix will urge the authorities to reassess the local safety system.*
- 2) *For the first time an authorised demonstration seemed to be out of their control.*
- 3) *This concise inventory has helped since 1982 the development of exposure schemes for the prevention of natural disasters.*
- 4) *Blood washed in Syria pn wednesday 10 February.*
- 5) *About 200 lodgings run the risks of lanslides, floods or avalanches.*

The above sentences will help us show the shift from natural language text to syntactic structures with representation of event type.

3.4.2. Presentation of formal concepts of the event notion

On analysing several sentences, we noticed that they may have one of the following forms :

- 1) **Occurence indicators** followed by an event. Example : *For the first time...*
- 2) Preposition followed by a **calendar term**. Example : *Since 1982...*

- 3) Event followed by a **calendar term**. Example : ... *on Wednesday 10 February*.
- 4) An event1 followed by a relative pronoun, followed by a **verb action**, followed by a **transitive verb**, followed by event2. Example : *The murderous avalanche **which hit** the valley of Chamonix, will urge the authorities **to re-assess***
- 5) Subject followed by a **transitive verb**, followed by event. Example : *About 200 lodgings **run the risk of** of lanslides.*

This representation has led us to draw the main **linguistic markers** and to sequence them according to their types :

The different classes are as follows :

1) The **calendar term** class

- a) **prp_num** stands for preposition + number. Example : *In 1998*
- b) **prp_cal_num** stands for : preposition + calendar + number. Example : *On October 1998.*
- c) **Prp_inf** stands for preposition + infinitive + preposition. Example : *From, starting from, to deduct of.*
- d) **Cal_num_cal** stands for calendar, number, calendar. Example : *wednesday 10 February.*
- e) **Ver_prt** stands for temporal preposition. Example : *comes after, occurs before, creates since.*

2) The **occurrence indicator** class

- a) **Adj_occ** stands for adjective + occurrence. Example : *another time, last time, first time.*
- b) **Adt_det_occ** stands for tense adverb + determiner + occurrence. Example: *once again.*

3) The **relative pronoun** class

- a) **Prr_aux_ppa** : relative pronoun + auxiliary + past participle. Example : *which hit.*
- b) **Prr_aux_adv_ppa** stands for relative pronoun + auxiliary + adverb + past participle. Example : *who drank too much.*

4) The **transitive verb** class

- a) **Aux_ppa_prp** stands for auxiliary + past participle + preposition. Example : *are exposed to, were loaded with, have led to.*

3.4.3. Extraction of factual sentences

The extraction process is based on the result of the morpho-syntactic analysis. These results, which are a translation of morpho-syntactic sentences, are skimmed in order to identify factual markers. We've decided to keep the sentence which presents one of the markers ; the latter being also sequences of morpho-syntactic categories.

In addition to these structure indicators (morpho-syntactic indicators), we added a list of verbs which illustrate some event classes as they are defined by Foucou (Foucou, 1998), examples : the class of natural catastrophy (take place) : floods, earthquakes landslides,...

The class of meteorological phenomena (occur) : fog, snow, storm, humicave, ...

This list will help us extract all factual sentences because we may find sentences which do not have any of the define markers that are based on the formal structure of the sentence (Cf Figure 2 and Figure 3).

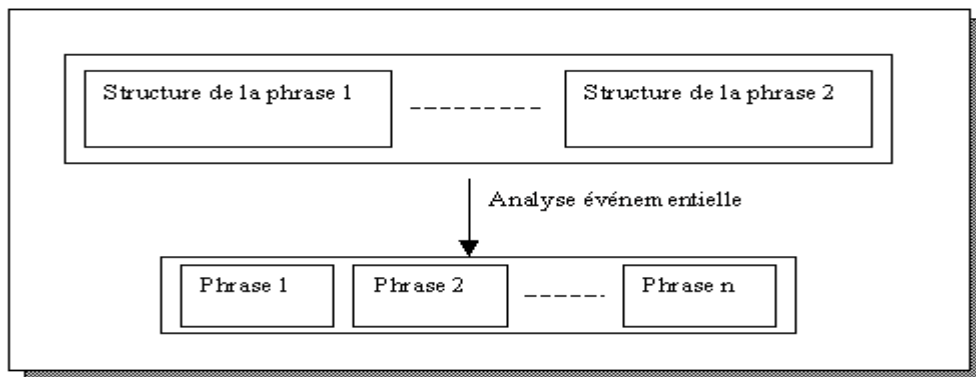


Figure 2: Descriptive diagram of the extraction module

The sentences in the above schema illustrate an example of sentences bearing factual information.

Structure :

1. prp_num_nr_aux_det_det_adj_nom_pro_pro_inf_prp_det_nom_adj_det_nom
2. prp_inf_prp_num_det_nom_ppa_aux_nr_prp_det_nom_det_nom_prp_nom_prp_nom_adj_adj_nr_ppa_prp_nom_adj

Texte :

1. En 1946, Cham onix fut l'une des toutes premières communes à se doter d'une carte locale des avalanches.
2. À partir de 1982, cet inventaire détaillé a servi à l'élaboration des plans d'exposition aux risques naturels prévisibles (PERN) réalisés au niveau communal.

Figure 3: Example of the extraction module

3.4.4. Interpretation of factual sentences

After extracting sentences which bear factual information , we will now try to answer a classical question according to the field of extraction but which is major importance and which is : who does what, to whom, when and where? (Cf. Figure 4).

The answer to the above question can be of great help especially if we want to extend our work and add a module called **text generation**.

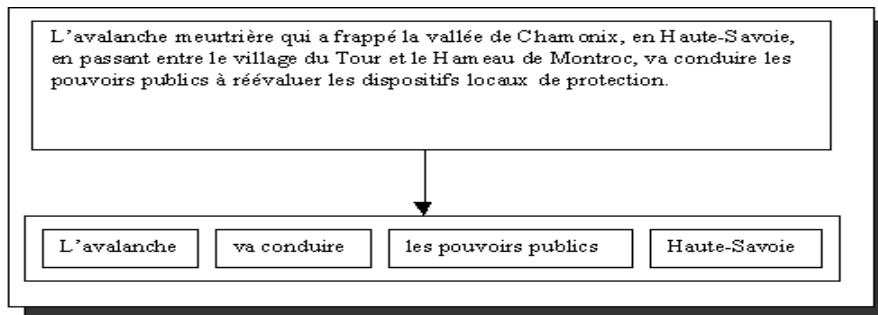


Figure 4: Example of the interpretation module

3.5. Enrichment of the list of markers

We also thought about a module of enrichment and consultation of the list of markers which have been defined.

It is quite useful to allow the user to add other markers or to define another list of markers so that each user of the system will be free to adapt it to his or her own needs.

This is not an easy task but it implies that the user knows perfectly well how the modules of the system function especially the morphological analysis module. Once the user inputs the markers he wishes to add, the system will suggest a morpho-syntactic structure.

4. Conclusion

Thanks to our morphological sensor based on inflectional morphology, we were able to directly extract type information as well as interpret the type of event ie, futur event or past event.

The EXEV system is conceived as a system based on linguistic knowledge and it has an interface of result consultation.

The system can be improved in two ways : we can on one hand increase the linguistic data base and on the other hand the interfacing of result consultation.

The validation of the EXEV system is carried out on technical texts namely of journalistic genre. Moreover, because its analysis modules and its chosen markers are independent from the documentary source, it allows its users to apply it on other types of texts such as Web documents or medical literature, etc. It also gives us, the possibility to extract other information relevant to other fields (other than event extraction) such as the causality notion. This can be done by inputting the markers related to the field, for example for the causality notion we must introduce in the basis the following markers : to result of, to be provoked by, to be due to, to cause, to provoke, ...

However, we believe that the markers are valid for all types of texts. In other words, the verbs chosen in our linguistic model which express the notion of event in French, can be found in any type of text with a semantic factual value.

The purpose of enriching the markers basis is to apply the EXEV system to other type of informtaion.

References

- Desclés J.P. (1993). L'exploration contextuelle : une méthode linguistique et informatique pour l'analyse automatique de texte, *ILN'93*, pp. 339-351, 1993.
- Berri J. (1995). SERAPHIN – Main Sentences automatic Extraction System. *Langage Engineering Convention*, London.
- Desclés J.P., Cartier E., Jackiewiz A., Minel J.L. (1997). Textual Processing and Contextual Exploration Method, *Proceedings of Context'97*, Universidade Federal do Rio de Janeiro, pp 189-197.
- Cartier E. (1998). Analyse automatique des textes : l'exemple des informations définitives. *Actes de la Rencontre Internationale sur l'extraction, le Filtrage et le Résumé automatiques (RIFRA'98)*, Sfax, Tunisie, 11-14 novembre, pp.6-18.
- Faïz R. (1998). Filtrage automatique de phrases temporelles d'un texte. *Actes de la Rencontre Internationale sur l'extraction, le Filtrage et le Résumé automatiques (RIFRA'98)*, Sfax, Tunisie, 11-14 novembre, pp.55-63.
- Faïz R. (1999). Automatic Extraction System of Textual Information. *Proceedings of the IASTED International Conference on Intelligent Systems and Control (ISC'99)*, Santa Barbara, Californie, Etats-Unis, 28-30 octobre, pp. 275-280.
- Foucou P. Y. (1998). Classes d'événements et synthèse de services Web d'actualité, *Actes de la Rencontre Internationale sur l'extraction, le Filtrage et le Résumé automatiques (RIFRA'98)*, Sfax, Tunisie, 11-14 novembre, pp.154-163.
- Garcia D., Aussenac-Gilles N., et Courcelle A. (1999). Exploitation, pour la modélisation des connaissances causales détectées par COATIS dans les textes. In *Ingénierie des connaissances* . Eds G. Kssel, J. Charlet et M. Zacklad. Editions Eyrolles, Paris ,pp. 44-54.
- Jacobs P. S. & Rau L. F. (1990). SCISOR : Extracting information from on-line news, *Commun. ACM* 33 (11), pp. 88-97.
- Teufel S., Moens M. (1998). Sentence extraction and rhetorical classification for flexible abstracts. *Proceedings of AAAI'98*, mars.