

Automatic Annotation Approach of Events in News Articles

Aymen Elkhilifi¹ and Rim Faiz²

¹LARODEC, ISG of Tunis
B.P.1088, 2000 Le Bardo, Tunisia
Aymen.Elkhilifi@isbb.rnu.tn
²LARODEC, IHEC of Carthage,
2016 Carthage Présidence, Tunisia
Rim.Faiz@ihec.rnu.tn

Abstract: *Daily, several news agencies publish thousands of articles concerning many events of all types (political, economic, cultural, etc.). The decision-makers find themselves in front of a great number of events, a few of which concern them. The automatic treatment of such events becomes increasingly necessary. Thus, we propose a machine learning-based approach that allows annotating news articles to generate an automatic summary of the events. We propose a new similarity measurement between events and we validate our approach by the development of the "AnnotEv" system.*

Keywords: *Document Annotation, Event Extraction, Machine Learning, Natural Language Processing.*

Received: August 30, 2008 | **Revised:** September 30, 2008 | **Accepted:** December 30, 2008

1. Introduction

Acquiring knowledge from texts is a need which has increased during the last years. With the considerable rise of the documents available in electronic format, it is necessary to extract, filter and analyse relevant information from those documents. As an example, the stock market events are numerous and diversified. The stock market experts must analyze these events, in a relatively reasonable time, in order to make important decisions. It is a question, therefore, of annotating the documents presenting the events to be able to extract those which are relevant. In this respect, our work aims at developing of an approach that annotates the news articles.

After the beginning of current Web's extension towards the semantic one, the annotation system starts to take a significant role. In fact, it participates in giving the semantic aspect to the different types of documents.

Furthermore, with the proliferation of news articles from different sources now available on the Web, summing up such information is becoming more and more indispensable. Due to the large number of news sources (such as: *BBC, Reuters, CNN, Aljazeera, etc.*), everyday, thousands of articles are produced in the entire world concerning a given event. That is why we

should think of automatizing the annotation process of such articles.

The documents indexing and events extraction are becoming tiresome. Thus, we are urged to generate an easily consultable semantic annotation that takes into consideration the increase in document size and enriches its indexing. By seeking a given event via a sequential analysis of the article, we noticed sentences which do not refer to any event. We observe, also, that several other sentences refer to the same event. That is why we intend to eliminate the non-event sentences, and to group the others in cluster form.

Accordingly, our work focuses on the annotation of documents: First, we prepare the text in a preprocessing stage. Second, we omit the non-event sentences. Then, we group the sentences indicating the same or similar events. Finally, we generate a summary article.

The rest of the document is organized as follows: Section (2) introduces the related work on annotation methods. Then, the particular methods of temporal information annotation are exhibited. In section (3), we present our approach for automatic events annotation. In order to validate our survey, we describe the four steps we followed to carry out the

AnnotEv system. The experimentation is described in section (4). In section (5) we evaluate the system in order to demonstrate its proficiency. In section (6) we end our work with a few notes on future work.

2. Related Works on Methods of Annotation

It is worth noting that the annotation definition varies according to the application domain: Linguistics [1, 2], E-learning [3], Biology [4], Software development [5] and Web application [6]. But, it can be said that the annotation is all graphic or textual information attached to a document. It refers to various entities including a set of documents, a document, a passage, a sentence, a term, a word or an image [7].

Several methods and techniques are used for the current annotation systems such as contextual exploration [8], conceptual graphs [9], meta-thesauri [4] and linguistic indicators [10]. We describe, thereafter, the main existing annotation systems.

SyDoM [9] is a semantic annotation system of Web pages. It allows the enrichment of these pages in order to find them without taking account of their writing language. It is devoted to the management of textual documents stored with XML formats. We see that SyDoM has two main advantages: first, multilingual research and second, the improvement of the Web pages representation. But, we notice that SyDoM can carry out research only on Web pages that it has already been annotated, yet it is unable to interrogate Web pages when the annotations were created using different semantic thesauri.

EXCOM [12] is an annotation engine that uses a set of linguistic tools, which aim at annotating a document by a bloc of internal/external knowledge. This engine is under development and, at the present time, it allows the production of a temporal organization of the stories and an automatic reformulation of the questions. However, we observe that an important part of this system is still not implemented, i.e. the semantic indexing of the documents by considering of annotated information.

Annotea [13] is a collaborative Customer Server system for document annotation. The annotations are stored on a specialized server. They are divided in such a way that anyone, who has an access to the annotation server will be able to consult all annotations related to a given document and to add his/her own annotations. These annotations can be typographical comments, corrections, assumptions or estimates. This system was developed using the W3C standards. Nevertheless, the only possible form of annotation is the text; it cannot annotate by images or icons.

The system developed by the ACACIA team [4] allows the annotation of genes. It helps the biologists, who make experiments on the biopuces, to validate and interpret of the obtained results. The system exempts them of the hard task of research. It offers the possibility of a keyword research in genetic databases. The keyword can correspond to genes or a biological phenomenon studied.

All previous works are interested in general documents annotation like scientific articles, Web documents, biological databases and multimedia documents. Only few of them focus on the events annotation. We present, in the following, some of these works:

The annotation of temporal information in texts [4]: this work focused more specifically on relations between events introduced by verbs in finite clauses. It proposes a procedure that achieves the task of annotation and a way of measuring the results. The authors of this work tested the feasibility of this procedure on newswire articles with promising results. Then, they developed two evaluation measures of the annotation: fineness and consistency.

The annotation scheme for annotating features and relations in texts [14]: it enables to determine the relative order and, if possible, the absolute time of the events. A scheme could be used to construct an annotated corpus. This corpus would yield the benefits normally associated with the construction of such resources. It can be also used to better understand the phenomena. Moreover, it represents a resource for training and evaluation of adaptive algorithms. Also it identifies automatically features and relations of interest. However, we noted that this work is based only on the temporal markers to determine the relations between events. This technique is not completely correct, since there are implicit inter-events relations which are expressed without using temporal markers.

The time annotation [15] with a canonized representation of the times expressions: a method was described for extracting such time expressions in multiple languages. The annotation process is divided into two steps: first, flagging a temporal expression in a document (based on the presence of specific lexical trigger words) and, second, identifying the time value that the expression designates or the speaker intends for it to designate.

We note that the temporal information annotations are generally concerned with the detection of dates and temporal markers [14], event descriptions and finding the events date [16] and the temporal relations between events in a text [11].

However, in our study we are interested rather in the annotation of the events in the form of metadata on the document.

3. The Proposed Approach of Event Annotation

We note that the above-mentioned approaches for the temporal information annotation are mainly linguistic. As well, they are based on the temporal indices. Moreover, we are interested in the annotation and exploitation of events using machine learning techniques. Our approach is not restricted to the events detection, but it also allows gathering the similar events in order to facilitate an ulterior treatment: indexing, storage in a database, summarization, categorization, information retrieval, etc.

The automatic process of documents annotation which we present is carried out in four stages (see Figure 1):

1. **Preprocessing:** it consists, on the one hand, in the text segmentation and, on the other hand, in the identification of entities.
2. **Events annotation:** it uses a classifier playing the role of a filter for the non-event sentences.
3. **Clustering:** it consists in gathering the sentences referring to the same or similar events. We propose in this stage a new similarity measurement between the events.

4. **Document annotation:** it takes various forms such as: sentences, form, concept, according to the field of application of our approach.

3.1 The Preprocessing

In the case study, the preprocessing consists in the application of some Natural Language Processing (NLP) tools to the rough text in order to segment it into sentences and annotate its entities.

We noticed that the segmentation is often neglected by the annotation systems in spite of its importance compared to the annotation quality. In addition, the entity identification is often used in other contexts like the question-answer systems. We present, thereafter, the segmentation and the named entity recognition.

The segmentation: is the determination of the sentences borders. 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 well structured texts into paragraphs. But, the segmentation of texts in smaller units (sentences) remains a more complex task.

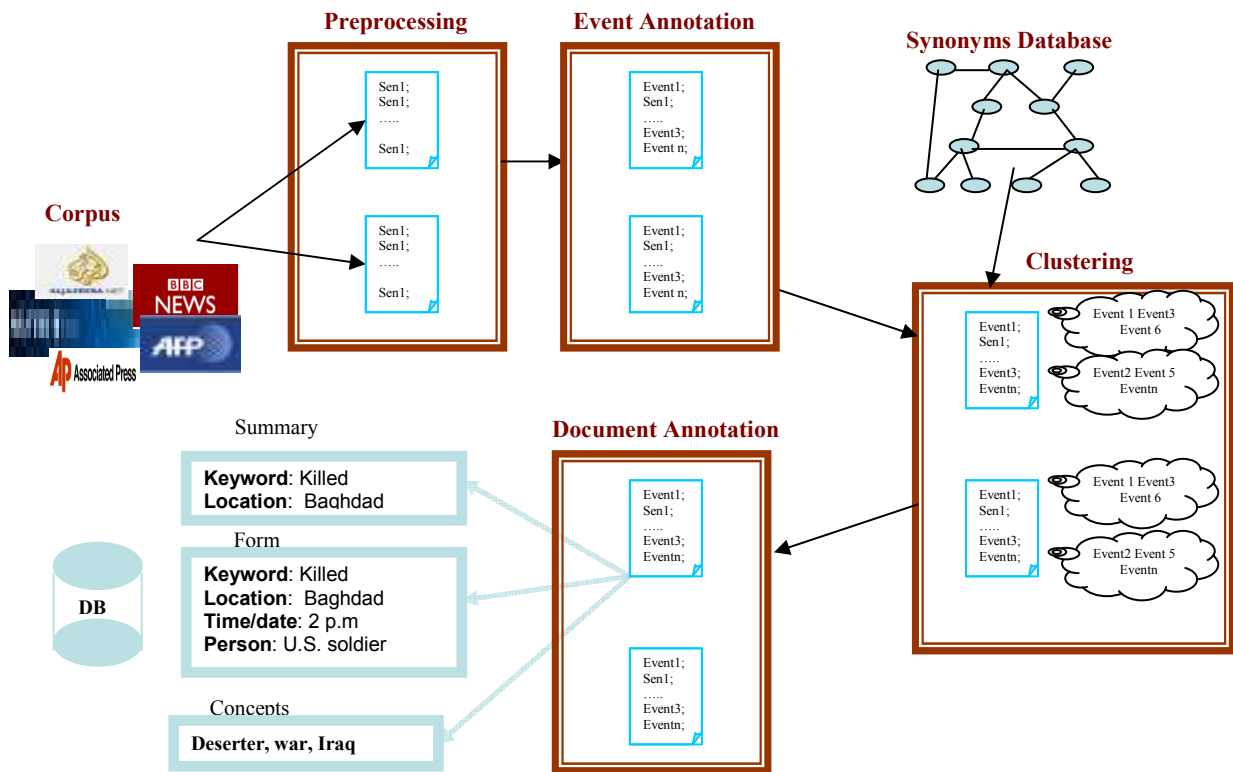


Figure 1. Proposed approach for automatic annotation of events

There exist some works related to the monolingual segmentation, in French language [17], English [18] and German [19]. We tested these systems on a corpus, and they gave a precision lower than 90%. Other more recent works considered the multilingual aspect, like the work of Mourad [20], which proposed an approach that consists in defining a textual segment starting from a systematic study of the punctuation marks.

We have developed our own segmentor while relying on punctuation marks. Due to the large number of linguistic rules to implement, we have to integrate in our knowledge base all the rules developed in the Segatex system [20]. We obtained 96% of precision. The result of the segmentation of a text is to detail below on a sample article (cf. Figure 2).

```
<?xml version="1.0" encoding="UTF-8"?>
<article lang="Ang">
  <para id="1">
    <title id="1">
      <phrase id="1">
        Iraqi leader denies civil war as 50 people die
      </phrase>
    </title>
  </para>
  <para id="1">
    <phrase id="2">
      BAGHDAD, Iraq (CNN) -- On a day in which at
      least 50 people were killed, Iraqi Prime Minister
      Nuri al- Maliki said he did not foresee a civil war
      in Iraq and that violence in his country was abating.
    </phrase>
  </para>
  <para id="3">
    <phrase id="3">
      <quote id="1">
        « In Iraq, we'll never be in civil war, »
      </quote>
      al-Maliki told CNN's « Late Edition » on Sunday.
    </phrase>
  </para>
  <para id="4">
    <phrase id="4">
      Attacks on American troops around the Iraqi
      capital Sunday left six soldiers dead, the U.S .
      command in Baghdad reported.
    </phrase>
  </para>
  <para id="5">
    <phrase id="5">
      Other violence nationwide left more than 130
      wounded, local authorities said.
    </phrase>
  </para>
  <para id="6">
    <phrase id="6">
      One U.S . soldier was killed by gunfire in eastern
      Baghdad about 2 p.m .
    </phrase>
  </para>
  <!-- ..... -
</article >
```

Figure 2. Extract of the article segmentation " *Iraqi leader denies civil war as 50 people die* ". BBC 2006.

Named Entities [21] are types of particular lexemes which refer to an entity of the concrete world in certain fields, namely human, social, political, economic or geographical and which has a name (typically a proper name or an acronym). The entities are identified in the documents by a tag which corresponds to the entity

type. The types selected are recognized by rules thanks to the joint exploitation of two information sources:

- General lexicons allowing finding syntactic and semantic features associated with the simple words in complement of the lexical features.
- Dictionaries of named entities.

It may be noted that the named entity recognition is an important task of text information extraction. The majority of the current systems are able to annotate the dates and the places. It is possible to find within one document several mentions which refer to only one entity.

Figure 3 presents the same text of Figure 2 after the named entities detection. Initially the position of each term is fixed. Then, the annotations concerning each entity are mentioned by specifying their attributes.

```
<?xml version="1.0" encoding="windows-1252" ?>
<!-- The document content area with serialized nodes -
<TextWithNodes>
  <Node id="4"/>Iraqi<Node id="9"/>
  <Node id="10"/>leader<Node id="16"/>
  <Node id="17"/>denies<Node id="23"/>
  <Node id="24"/>civil<Node id="29"/>
  <Node id="30"/>war<Node id="33"/>
  <Node id="34"/>as<Node id="36"/>
  <Node id="37"/>50<Node id="39"/>
  <Node id="40"/>people<Node id="46"/>
  <Node id="47"/>die<Node id="50"/>
  <Node id="56"/>BAGHDAD <Node id="63"/>
  , <Node id="64"/>
  <!-- ..... -
</TextWithNodes>
<AnnotationSet>
  <Annotation Id="485" Type="Location" StartNode="56"
  EndNode ="63">
    <Feature>
      <Name
      className="java.lang.String">rule2</Name>
      <Value
      className="java.lang.String">LocFinal</Value>
    </Feature>
    <Feature>
      <Name
      className="java.lang.String">rule1</Name>
      <Value
      className="java.lang.String">Location1</Value>
    </Feature>
    <Feature>
      <Name
      className="java.lang.String">locType</Name>
      <Value
      className="java.lang.String">city</Value>
    </Feature>
  </Annotation>
  <Annotation Id="478" Type="Person" StartNode="305"
  EndNode ="314">
    <Feature>
      <Name
      className="java.lang.String">gender</Name>
      <Value
      className="java.lang.String">male</Value>
    </Feature>
    <Feature>
      <Name
      className="java.lang.String">rule1</Name>
      <Value
      className="java.lang.String">PersonFull</Value>
    </Feature>
```

```

<Feature>
  <Name
  className="java.lang.String">rule</Name>
  <Value
  className="java.lang.String">PersonFinal</Value>
</Feature>
<Feature>
  <Name
  className="java.lang.String">matches</Name>
  <Value className="java.util.ArrayList"
  itemClassName =
  "java.lang.Integer">476;478</Value>
</Feature>
</Annotation>
<Annotation Id="431" Type="Split" StartNode="561"
EndNode="562">
  <Feature>
    <Name
    className="java.lang.String">kind</Name>
    <Value
    className="java.lang.String">internal</Value>
  </Feature>
</Annotation> <!-- ..... -
</AnnotationSet>

```

Figure 3. Extract article "Iraqi leader denies civil war as 50 people die" after the named entity recognition.

We integrate in this module the dictionaries of GATE software [34] (lists of the entities and regular expressions). The result of this first stage of our approach is the set of segmented sentences with their annotated entities.

3.2 Events Annotation

An event is a specific object which occurs at one specific moment and in a well defined place [22]. Our objective is to identify all events in a document. We mark each detected event by a tag. Accordingly, a model of classification is built automatically from the training set which permits to predict whether a sentence contains an event or not. We initially used the attributes which refer to the events as they are defined by Naughton [23]. These attributes are the following:

- Length of the sentence.
- Numbers of capital letters.
- Numbers of stop words.
- Number of city/town.
- Number of numerical marks.

Within the framework of our study, and through the analysis of the news articles, we noticed that the addition of other attributes to the preceding list is possible; e.g. the temporal markers (*after, before, simultaneously, etc.*) and the calendar terms (*Sunday, 9/12/2004, March*). The problem of choosing significant attributes can be solved by using feature selection algorithms, which lead to select a subset of relevant attributes in order to find a predictive model. There is a variety of feature selection algorithms (chi-public garden, Relief and Principal component analysis [24]). After having carried out experimentation, we added the attribute "number of calendar terms".

Several machine learning techniques can be used for classification problem such as the neural network, the

decision tree, the Bayesian network, etc. We chose the decision tree for many reasons; it is easily interpretable by people. Moreover, the decision tree construction is less skeletal compared to the other techniques. Hence, it allows the reduction of the system complexity.

The training set is annotated by experts. For each news article, the events are annotated as follows: the annotator is brought to assign labels for each sentence representing an event. If a sentence refers to an event, they assign the label "yes", if not, "no". We applied to this same training set various algorithms of decision trees construction. Then, we chose the model which has the biggest *PCC* (Percentage Correctly Classified).

The result of this stage is the set of sentences referring to events. Moreover, the classification of sentences as an event or not, represents a kind of filtering; on the basis of a segmented text, we filter the non-event sentences.

3.3 Clustering

In this stage we gather the sentences referring to the same or similar events by the application of the algorithm 'Hierarchical Agglomerative Clustering (HAC)' [25, 26]. 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 is to put forward a new measure of similarity between events. Given the importance of similarity measurements in clustering, we noted that there are several of such measurements between documents including: Salton's cosine [27], Khi-Deux distance [28], Cosine in distributional space [29]. Other measurements, which are more interesting for us, are linked to the similarity between sentences, the latest of which is Naughton's measurement [23].

Similarity between sentences: Similarity measurement, in general, is based on distance (Euclidean, Manhattan, Minkowski or that of Entropy [30]). We can adopt the Jaccard index for sentences. If we replace a document by a sentence in the formula, we get: $S_{ij} = m_c / (m_i + m_j - m_c)$.

The similarity index is the number of common words divided by the total number of words minus the number of common words:

- i and j are two sentences
- m_c : Number of common words.
- m_i : Size of the lexicon of the sentence S_i (i.e. number of different words in S_i).
- m_j : Size of the lexicon of document S_j .

We put forward a new similarity measurement between events inspired by *tf-idf* "weight term frequency-inverse document frequency" [31]. This

measurement also takes account of the clusters position in the article.

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

C_2 : *In other attacks reported by security and hospital officials, two car bombings in the northern city of Kirkuk killed 10 and wounded 32, and a blast in the southern 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:

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

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:

$$Ct_{ij} = tf(t_i, c) \times \log(N/df(t_i)) \text{ with:}$$

- $tf(t_i, c)$ 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 .

Based on what has been said so far, and by taking into consideration the sentence position in the article, we propose the new similarity measurement **FSIM** which combines the similarity between sentences 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 $\alpha \in [0, 1]$ fixed during the experimentation. Therefore, for N clusters, we have $n \times (n-1)/2$ possible combinations (see [24]).

It is important to group the sentences indicating the same or similar events, since they will be gathered even if they use various words. Figure 4, for example, presents the application of HAC algorithm by using

FSIM on a press article:

C₁: Iraqi leader denies civil war as 50 people die.

C₂: On a day in which at least 50 people were killed, Iraqi Prime Minister Nuri al-Maliki said he did not foresee a civil war in Iraq and that violence in his country was abating.

In Iraq, we'll never be in civil war," al-Maliki told CNN's "Late Edition" on Sunday.

C₃: One U.S. soldier was killed by gunfire in eastern Baghdad about 2 p.m.

C₄: U.S. commander wounded since 1 p.m

.....

Levin, the ranking Democrat on the Senate Armed Services Committee, called for the United States to set a date to begin withdrawing its forces.

C₅: In Baquba, two separate shooting incidents Sunday afternoon left six dead and 15 wounded, officials said.

C₆: Two car bombings in the northern city of Kirkuk killed 10 and wounded 32, and a blast in the southern city of Basra killed five and injured 15

FSIM (C_1, C_2) = **0.91** FSIM (C_1, C_3) = **0.12**

FSIM (C_1, C_4) = **0.1** FSIM (C_1, C_5) = **0.05**

FSIM (C_1, C_6) = **0.02** FSIM (C_2, C_3) = **0.08**

FSIM (C_2, C_4) = **0.1** FSIM (C_2, C_5) = **0.32**

FSIM (C_2, C_6) = **0.36** FSIM (C_3, C_4) = **0.84**

FSIM (C_3, C_5) = **0.28** FSIM (C_3, C_6) = **0.23**

FSIM (C_4, C_5) = **0.19** FSIM (C_4, C_6) = **0.15**

FSIM (C_5, C_6) = **0.79**

Figure 4. Application of HAC algorithm (first step)

The sentences in bold indicate an event. First of all, we calculate the FSIM between these sentences where each sentence is a cluster. Then, we obtain values at the bottom of figure 4. Besides, we group together C_1 and C_2 into only one cluster C_A (because they have the biggest FSIM). Finally, we reapply HAC on the new clusters.

After 3 iterations, we obtain two new clusters C_B and C_C (cf. Figure 5):

C_A: Iraqi leader denies civil war as 50 people die.

On a day in which at least 50 people were killed, Iraqi Prime Minister Nuri al-Maliki said he did not foresee a civil war in Iraq and that violence in his country was abating.

In Iraq, we'll never be in civil war," al-Maliki told CNN's "Late Edition" on Sunday.

C_B: One U.S. soldier was killed by gunfire in eastern Baghdad about 2 p.m.

U.S. commander wounded since 1 p.m

.....

Levin, the ranking Democrat on the Senate Armed Services Committee, called for the United States to set a date to begin withdrawing its forces.

In Baquba, two separate shooting incidents Sunday

C_C: afternoon left six dead and 15 wounded, officials said.

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

FSIM (C_A, C_B) = **0.14**

FSIM (C_A, C_C) = **0.07**

FSIM (C_B, C_C) = **0.09**

Figure 5. Application of HAC algorithm (last step)

In this HAC step, we stop the clustering since we have a value of *FSIM* inferior to the threshold of similarity, fixed initially at 0,6.

3.4 Document Annotation

Using the clusters and their positions in the article we generate a description which combines the events and presents the annotation of the article under three types:

- To extract sentences that sum up the article.
- To structure the annotation in a standard form to store events in databases.
- To extract concepts (future work).

Thus, we continue the enrichment of the document by other metadata which will be very useful for all ulterior treatment (information retrieval, automatic summarization, question-answer systems, storage of the events in a database, indexation, etc).

A possible form of metadata is the forms filling, i.e. stoking the events in database while answering has well determined questions (where, when, who) for example:

- Location: Baghdad.
- Time/date: 2 p.m.
- Person: U.S. soldier.
- Keyword: Killed.

For each event sentence, we have this information since the preprocessing. After having stored this information in a relational database, we can find events by date, person, or time by a simple request on the selected fields.

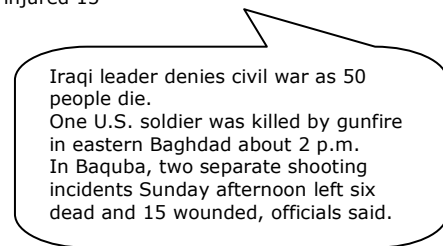
Another form of metadata is the automatic summary, which consists in marking the sentences which form the summary of a document. In general, the goal of a summary system is to produce a condensed representation of the contents where the important information of the original text is preserved. It is also necessary to consider that the user needs and the specified task [32].

In this context, we propose an informative summary containing the essential information of the article. This summary is also selective since it neglects the general aspects of the article. In addition, it can be said that it is targeted since it is correlated to the events.

For each cluster generated by the third stage we annotate the article by the principal events it contains. We use the following heuristics: the sentence having the maximum value attributes in the classification stage is the best to annotate the cluster. Let us take again the preceding example, the summary of which is presented as follows (cf. Figure 6):

Iraqi leader denies civil war as 50 people die.
 On a day in which at least 50 people were killed, Iraqi Prime Minister Nuri al-Maliki said he did not foresee a civil war in Iraq and that violence in his country was abating.
 In Iraq, we'll never be in civil war," al-Maliki told CNN's "Late Edition" on Sunday.
 One U.S. soldier was killed by gunfire in eastern Baghdad about 2 p.m.
 U.S. commander wounded since 1 p.m.

 Levin, the ranking Democrat on the Senate Armed Services Committee, called for the United States to set a date to begin withdrawing its forces.
 In Baquba, two separate shooting incidents Sunday afternoon left six dead and 15 wounded, officials said.
 Two car bombings in the northern city of Kirkuk killed 10 and wounded 32, and a blast in the southern city of Basra killed five and injured 15



Iraqi leader denies civil war as 50 people die.
 One U.S. soldier was killed by gunfire in eastern Baghdad about 2 p.m.
 In Baquba, two separate shooting incidents Sunday afternoon left six dead and 15 wounded, officials said.

Figure 6. Summarization of article

Such annotation can be used to improve indexing and information relevant to such articles.

4. Experimentation

To validate our approach, we develop the AnnotEv system with Java language; Visual Studio platform, particularly Visual J++.

AnnotEv includes the four following modules:

- Module 1: The segmentation and the recognition of the named entities
- Module 2: events annotation.
- Module 3: clustering.
- Module 4: automatic summary.

We prepared a corpus containing 263 articles. The articles talk about the Iraq War. They are published in 2008 and collected from 8 news sources (agencies and channels): CNN (32 articles), Reuters (30 articles), BBC (35 articles), Associated Press (31 articles), AFP (35 articles). Aljazeera (32 articles), Iraqi Press Agency (33 articles) and Iran Islamic Republic press agency (35 articles).

The average length of a sentence is of 14 words, with an average of 7 events per article, for a total of approximately 281000 words, 20071 sentences and 1841 events.

We also use the Weka libraries edition 3.5 [33] and Gate 4.0 [34] respectively for decision tree and entity identification.

After removing the images and the legends of the article we segment them into sentences and, then, we annotate entities by calling upon the Gazettee-Annie method from Gate class. We use WordNet [35] as a synonyms database. Besides, we annotate the events and group them according to their similarities. We develop several interfaces to ensure the management of corpus and training set annotation. We also develop other interfaces for the events search and summarization. We have to integrate a speaker agent who is able to read the texts in two different versions. For the system user, it is enough to select a new article in order to listen to its summary (see Figure 7).

The Precision, the Recall and F1 are defined as follows:

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

5. Results

The evaluation is done at several levels. We start with the classification evaluation by using the PCC, then, the clustering by measuring the Precision and the Recall. We exploit the following algorithms:

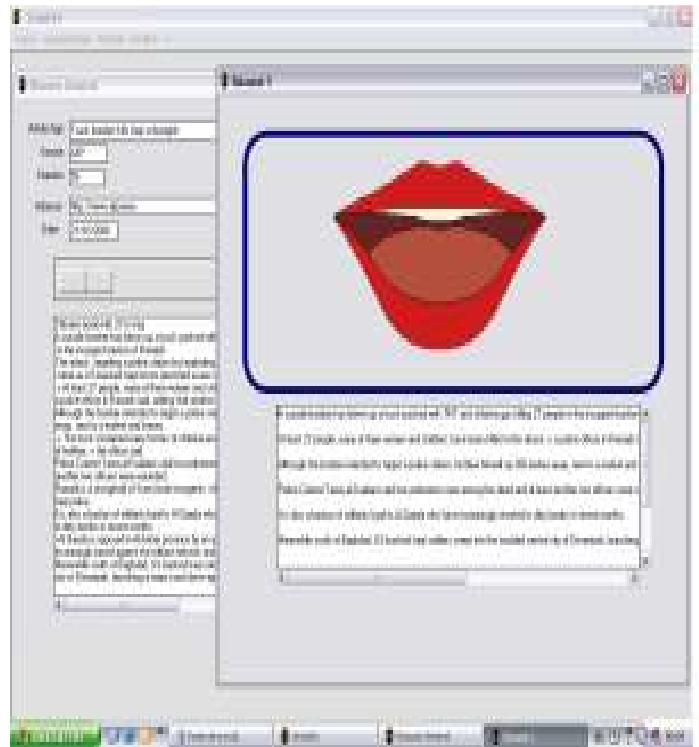


Figure 7. Interfaces of speaker agents for the automatic summarization

The training set is part of the group of obtained sentences after the preprocessing stage. It is annotated by two experts. For each sentence the default value of the attribute “Event” is 'No' (sentence not indicating an event), the commentator has to put 'Yes' if the sentence refers to an event. An ARFF file (input format of Weka) is generated automatically for each article. It will be used like a data source for the algorithms of classification. We adopted J48, ADTREE and Random Tree with the cases of the events.

To evaluate the clustering method we employ the precision and recall measurements [35]. We assign each pair of sentences to 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.

J48: implementation of C4.5 algorithm [36] which selects for each level the tree node as the attribute which differentiate better the data. Then, it divides the training set into sub-groups in order to reflect the values of the attribute of the selected node. We repeat the same treatment for under group until we obtain under homogeneous groups (all the instances or the majority have the same attribute of decision).

ADTree: construction of the decision trees extended to the cases of multiclass and multi-labels.

Random Tree: begin with tree random and chosen by the majority best vote.

We obtained the following results:

J48

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.824	0.25	0.778	0.824	0.8	Yes
0.75	0.176	0.8	0.75	0.774	No

ADTREE

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.765	0.313	0.722	0.765	0.743	yes
0.688	0.235	0.733	0.688	0.71	no

RandomTree

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.706	0.313	0.706	0.706	0.706	yes
0.688	0.294	0.688	0.688	0.688	no

For the clustering stage, we obtained a significant value of Recall (R) and Precision (P) and the function F1:

- R = 81,43%
- P = 79,12%
- F1 = 80.25%.

This result is made of the similarity measurement that we proposed. Indeed, it detects the similarity between the sentences even if it contains different terms.

6. Conclusion and Future Work

In this paper, we have proposed four stages to annotate press articles starting, in a first stage, by the preprocessing that consists in applying NLP tools to prepare the data. In a second stage, the filtering of the non event-driven sentences has been done by dint of a classifier. In the third stage we have gathered the sentences in clusters according to their degree of similarity (*FSIM*). Finally, we have generated an automatic summary of the principal events constituting the article. Our approach was evaluated on news articles corpus concerning the Iraq War published in 2008.

This approach comes within the framework of the Information Extraction from texts, particularly the extraction and the exploitation of the events. Actually, it constitutes a considerable target in many application domains like the national security, the economy or the industry. In such fields the concepts of technological/economic survey become essential, in particular for the help in decision making: definition of strategies, placement towards competition, etc.

In short term, in our future works, we propose to adopt *AnnotEv* for news articles in Arabic. Indeed, our system supports the Arab characters, and the stages of our approach are independent of the language. But, it remains to provide linguistic resources (Synonyms data base, segmentators, etc) for the Arabic language. Experimentation was made on a small corpus of 5 articles segmented manually and with a small Synonyms data base to measure.

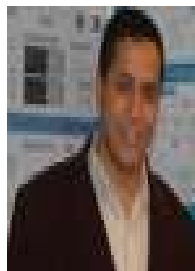
In long term, we look forward to fuse the events. In fact, we have the idea of adopting, to the case of the events, the MCT model proposed by Smets [37] for the fusion of information in general.

References

- [1] J. Véronis, *Annotation automatique de corpus : panorama et état de la technique*. J.-M. Pierrel (Ed.), Ingénierie des langues. Paris: Editions Hermès, 2000.
- [2] S. Bird, M. Liberman, A formal framework for linguistic annotation, in *Speech Communication*, Volume 33, Number 1, January 2001, pp. 23-60(38) [http://dx.doi.org/10.1016/S0167-6393\(00\)00068-6](http://dx.doi.org/10.1016/S0167-6393(00)00068-6)
- [3] M. Dominique, Modèles et outils logiciels pour l'annotation sémantique de documents pédagogique. *Thèse de doctorat de l'Université Joseph Fourier de Grenoble*, octobre 2006.
- [4] K. Khelif, R. Dieng-Kuntz, P. Barbry, An Ontology-based Approach to Support Text Mining and Information Retrieval in the Biological Domain, in *J. UCS* 13(12), pp. 1881-1907, 2007.
- [5] M. Islam Chisty, An Introduction to Java Annotations, 2005 at http://www.developer.com/java/other/article.php/10936_3556176_1
- [6] L. Denoue, L. Vignollet, An annotation tool for Web browsers and its applications to information retrieval, in *Proceedings of RIA*, 2000.
- [7] E. Desmontils, C. Jacquin, Annotations sur le Web: notes de lecture. In *AS CNRS Web Sémantique*, 2002.
- [8] J.P. Desclés, Systèmes d'exploration contextuelle, Co-texte et calcul du sens, Claude Guimier, *Presses de l'université de Caen*, 1997, pp. 215-232.
- [9] C. Roussey, S. Calabretto, An experiment using Conceptual Graph Structure for a Multilingual Information System, in *the 13th International Conference on Conceptual Structures*, ICCS'2005.
- [10] F. Dau, M-L Mugnier, G. Stumme, Conceptual Structures: Common Semantics for Sharing Knowledge, *13th International Conference on Conceptual Structures*, ICCS 2005.

- [11] P. Muller, X. Tannier, Annotating and measuring temporal relations in texts. In *Proceedings of Coling, volume I. Genève*, Association for Computational Linguistics, 2004.
- [12] J.P. Desclés, B. Djioa, La recherche d'informations par accès aux contenus sémantiques : vers une nouvelle classe de Systèmes de Recherches d'Informations et de moteurs de recherche (aspects linguistiques et stratégiques). *Rapport Paris Sorbonne et Revue Linguistique informatique, Roumanie*. pp. 7-57, 2007.
- [13] J. Kahan, M-R. Koivunen, Annotea: an open RDF infrastructure for shared Web annotations. *Proceedings of the 10th international conference on World Wide Web*, 2001.
- [14] A. Setzer, R. Gaizauskas, TimeML: Robust specification of event and temporal expressions in text. In *The second international conference on language resources and evaluation*, 2000.
- [15] I. Mani, L. Ferro, B. Sundheim, G. Wilson, Guidelines for Annotating Temporal Information. In *Human Language Technology Conference*. San Diego, California, 2001.
- [16] Faiz, R. Identifying relevant sentences in news articles for event information extraction. *International Journal of Computer Processing of Oriental Languages (IJCPOL)*, World Scientific, Vol. 19, No. 1, pp. 1-19, 2006.
- [17] A. Dister, Problématique des fins de phrase en traitement automatique du français. In *À qui appartient la punctuation ? Actes du colloque international et interdisciplinaire de Liège*, Bruxelles, 1997.
- [18] C. Jeffrey, C. Reynar, A. Ratnaparkhi, A Maximum Entropy Approach to Identifying Sentence Boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D.C., March 31 - April 3 1997.
- [19] D. Palmer, A. Hearst, Adaptive sentence boundary disambiguation. *Proceedings of the 1994 Conference on Applied Natural Language Processing*. Stuttgart, Germany, 1994.
- [20] Gh. Mourad, Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique de citations. Réalisation des Applications informatiques: SegATex et CitaRE. *Thèse de doctorat Université Paris-Sorbonne*, 2001.
- [21] O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, C. Jacquemin, L. Monceaux, I. Robba, A. Vilnat, How NLP Can Improve Question Answering. In *Revue Knowledge Organization*, 2002.
- [22] W. Li, X. Wang, A. McCallum. In *Event Extraction and Synthesis Conference Menlo Park*, California, 2006, pp. 48-54.
- [23] M. Naughton, N. Kushmerick, and J. Carthy, Event extraction from heterogeneous news sources. *Proc. Workshop Event Extraction and Synthesis, American Nat. Conf. Artificial Intelligence*, 2006.
- [24] A. Elkhilfi, and R. Faiz, Machine Learning Approach for the Automatic Annotation of Events. *Proceedings of the 20th International FLAIRS 2007 - Special Track: Automatic Annotation and Information Retrieval: New Perspectives*, D. Wilson and G. Sutcliffe (Editors), AAAI Press, California, pp. 362-367.
- [25] H. Liu, L. Yu, Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Transaction on Knowledge and Data Engineering*, 2005. vol. 17, n. 3.
- [26] Mannig, C., and Schutze, H. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
- [27] G. Salton, C. Buckley: Term-weighting approaches in automatic text retrieval. In *Information Processing & Management*, 24(5), pp.513-523, 1988.
- [28] L. Lebart, M. Rajaman, Computing Similarities. In Dale R., Moisl H. Somers. Editors: *Handbook of Natural Language Processing*, Marcel Dekker. New York, 2000.
- [29] B. Escofier, Analyse factorielle et distances répondant au principe d'équivalence distributionnelle, *Revue de Statist. Appl*, vol. 26, n°4, 29-37, 1978.
- [30] F. Alter, Y. Matsushita, X. Tang, An Intensity Similarity Measure in Low-Light Conditions, *ECCV06 (IV: 267-280)*.
- [31] G. Salton, M.J. Mac Gill, Introduction to Modern Information Retrieval. In *International Student Edition*, 1983.
- [32] J.L. Minel, J.P. Desclés, E. Cartier, G. Crispino, S. Ben hazez, A. Jackiewicz, Résumé automatique par filtrage sémantique d'informations dans des textes. *Revue Technique et Science Informatiques*, 2001.

- [33] E. Frank, I. Witten, *Data Mining: Practical Machine Learning Tools and Techniques* (Second Edition). Morgan Kaufmann, 2005. 525 pages.
- [34] K. Bontcheva, V. Tablan, D. Maynard, H. Cunningham, Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*. 10 (3/4), pp. 349-373, 2004.
- [35] C. Fellbaum, J. Grabowski, S. Landes, Performance and confidence in a semantic annotation task, In C. FELLBAUM, Ed., *WordNet: an electronic lexical database, Language, Speech and Communication*, chapter 9, pp. 216-237. Cambridge, Massachusetts: The MIT Press, 1998.
- [36] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc, 1993.
- [37] Ph. Smets, The Transferable Belief Model and other Interpretations of Dempster-Shafer's Model. *Uncertainty in Artificial Intelligence 6*, P.P. Bonissone, M. Henrion, L.N. Kanal, J.F. Lemmer (Editors), Elsevier Science Publishers, 1991, pp. 375-383.



Aymen Elkhilfi received a Masters degree in computer sciences in 2007 from the University of Tunis, High Institute of Management. He is carrying out a PhD degree in Computer Sciences between University of Tunis and Sorbonne University (Paris IV), France.

His research is focused on document annotation with the Machine Learning techniques and their applications on the semantic Web.



Prof. Dr. Rim Faiz obtained his Ph.D. in Computer Science from the University of Paris-Dauphine, in France. She is currently a Professor of Computer Science at the Institute of High Business Study (IHEC) at Carthage, in Tunisia. Her research interests include Artificial

Intelligence, Natural Language Processing, and Web Mining. Dr. Faiz is also the responsible of the MBA of New Technologies and Electronic Commerce at the Institute of High Business Study.