Indexing Documents by Discourse and Semantic Contents from Automatic Annotations of Texts

Brahim Djioua
LaLICC - Sorbonne University
28 rue Serpente 75006 Paris- France
bdjioua@paris4.sorbonne.fr

Jean-Pierre Desclés
LaLICC - Sorbonne University
28 rue Serpente 75006 Paris- France
Jean-Pierre.Descles@paris4.sorbonne.fr

Abstract

The basic aim of the model proposed here is to automatically build semantic *metatext structure* for texts that would allow us to search and extract discourse and semantic information from texts indexed in that way. This model is built up from two engines: The first engine, called EXCOM (Djioua et al., 2006), is an XML based system for an automatic annotation of texts according to discourse and semantic categories. The second engine called MOCXE uses automatic semantic annotation that is generated by EXCOM to create a semantic inverted index which is able to find relevant documents for queries associated with discursive and semantic categories such as *definition*, *quotation*, *causality*, *relations between concepts*, etc. We explain by an example of a relation of *"connection"* between concepts in French. The model used is enough general to be translated in other languages.

General presentation

Current existing web search engine systems that index texts generate representations as a set of simple and complex index terms, so that a classic search engine perform this model for information retrieval. It answers to queries defined in the form of linguistic terms which could be connected by logical operators such as AND, OR, NOT, etc. The answers are given by a list of documents, classified in a certain order. Certain sentences are used to illustrate answers corresponding to the initial query. The classification order is calculated with mathematical formulas according to the term frequencies composing the request and other factors not always revealed. In the existing web search tool, it is assumed that the major problem in current retrieval systems is to capture the meaning that a document may have for its users. This article explains how a new kind of web search engine is implemented by using semantic and discourse automatic annotation.

This paper describes two dependant automatic engines for semantic annotation and indexation based on linguistic knowledge. We'll try to explain how linguistic information (especially the discourse and semantic organization of texts) helps to retrieve relevant information, better than with a traditional web search engine.

For example, let us examine a typical answer from existing web search engine for the following question :

(1) : "Who met Chirac in December 2005 ?"

To ask a web search engine like Google for this question, we have to write the following query :

```
``Chirac AND met AND December AND
2005"
```

More precisely we must define the date range in the Julian Calendar for the month of December, in year 2005. Let us see just the relation of meeting between *"Chirac"* and other named-entities. The general search processing for Google is to associate the query terms in a relation of co-occurrence. The famous web search engine gives answers with documents which contains the linguistic terms *"Chirac"* and the verb *"met"*. That documents which contains *"Chirac"* in the beginning of the document and *"met"* at the end are selected. Of course, the user can use the proximity operator to constraints the term positions like `''Chirac * * * met"`, but we can not move away documents with the first term located at the end of a sentence and the second term located at the beginning of the next sentence.

Furthermore, a classic search engine can not select documents which contains the noun *"Chirac"* with the verbs *"visit"* or *"interviewed"* because of its indexation model which deals with terms rather than semantic categories.

In the context of the semantic web, electronic documents are marked up with metadata, using manual or semi-automatic annotation with web-based knowledge representation languages such as RDF and OWL (Berners-Lee, 2001) for describing the content of a document. The aim of this work is to encourage the automatic annotation of electronic documents and to promote the development of annotation-aware applications such as content-based information presentation and retrieval. Natural language applications, such as information extraction and machine translation, require a certain level of semantic analysis, which in practical terms means the annotation of each content segment with a semantic category for discourse (e.g, definition, causality, quotation or relation between named-entities).

We have noticed, in the last decade, at least two re-

search orientations in web information retrieval. One of two is qualified by "linguistic extension" which uses morphological flexion and syntactic schemas to index a "group of terms" - a terminology - rather than just terms founded in documents. An effort is made to find the best way to assign the right weight to documents for their relevant value with the request terms. The other orientation is the Berners-Lee "Web Semantic" which deals with manual or semi-automatic annotations based on domain ontologies. An other approach, that we support, is to use discursive organizations of natural language texts to define an other kind of information retrieval system. This model is not in opposition with the two previous approaches, but it is made as a complement system which is able to use morpho-syntactical extensions for terms, marked up named-entities (e.g, proper names, locations, time expressions, etc.) and relation terms in a domain terminology and also, in special contexts, can help to generate parts of domain ontology automatically from texts.

Related work: KIM plateform

The KIM platform (Kiryakov, 2004) provides infrastructure and services for automatic semantic annotation, indexing, and retrieval of documents. It allows scalable and customizable ontology-based information extraction (IE) as well as annotation and document management, based on GATE plateform (Cunningham, 2002). The model used in KIM believes that massive automatic semantic annotation is the prerequisite for the build-up of most of the metadata, needed for the Semantic Web. For each entity, mentioned in the text, KIM provides references (URI) (i) to the most relevant class in the ontology, and (ii) to the specific instance in the knowledge base. As a result of the automatic semantic annotation, metadata is generated and associated with the resource processed. This metadata is not embedded in the processed document, thus allowing different semantic annotation tasks to take place, accordingly resulting in diverse sets of metadata.

The couple GATE-KIM is focused to build automatically instances of an ontology by named-entities extraction using the pattern language Jape. The interconnected engines EXCOM-MOCXE are used to, first, automatically annotate textual segments with discourse and semantic organizations by using Contextual Exploration method and in a second time index these textual segment within their semantic annotations to provide a new kind of information retrieval by given to a user to perform queries with discourse and semantic categories.

To illustrate this difference between the two models used by KIM-GATE and EXCOM-MOCXE, let's use the following example :

*(2) Gordon Brown **met** George Bush during his two day visit.*

*(3) Ten days ago, when Blair **was interviewed by** the BBC's Jeremy Paxman, the prime minister was asked repeatedly whether he had seen that advice.*

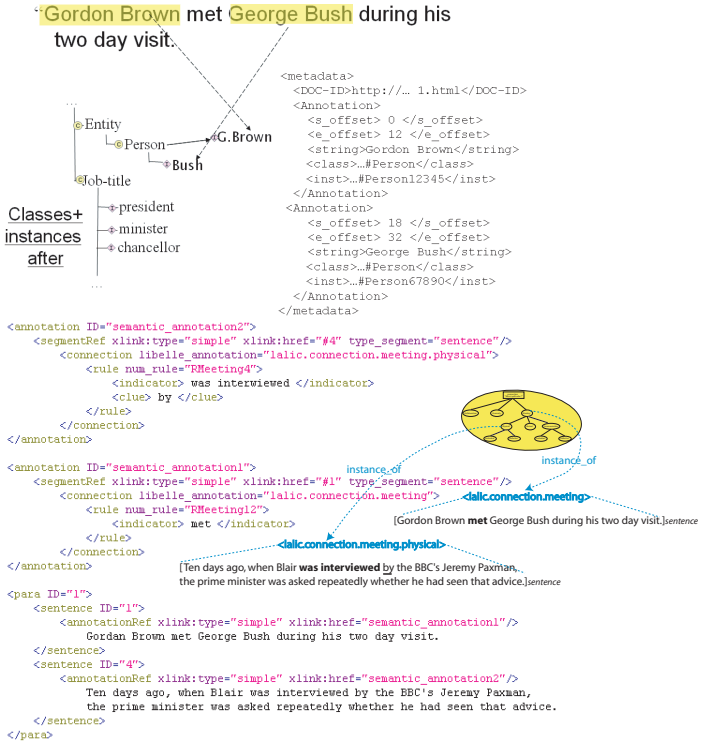

Figure 1: Example of an annotated text with GATE and EXCOM

Semantic "point of view" for text mining

A user's search relevant information proceeds by guided readings which gives preferential processing to certain textual segments (sentences or paragraphs).

The aim of this hypothesis is to reproduce: "what makes naturally a human reader" who underlines certain segments relating to a particular point of view which focuses his/her attention. There are several points of view for text mining of discourse organizations.

Indeed, such a user could be interested by the identification of the relations of *causality* by formulating a request such as: *find documents which contains "the causes of the tsunami"*. Another user will search by exploring many texts (specialized encyclopedias, handbooks, articles) the *definitions* of a concept (for example "social class" in sociology , "inflation" in economy , "polysemic" in linguistics,...).Yet an other user may be interested, by consulting the past five years old press, to know *connections* and the *meetings* which could take place between two named-entities (for example *"Poutine"* and *"Chirac"*, or between "the American Secretary of State" and "the Vice President of the United States"). Another will seek to establish connections, possibly by transitivity, between several suspects which however seem not to know each other, i.e. by establishing that "A" knows "B", had a phone call with him (according to a document of police force) and that "B" knows "C", had a lunch together (according to another document of police force).

The aim of these points of view for text mining is at a focusing reading and a possible annotation of the textual

segments which correspond to a research guided in order to extract information from them. Each point of view, as we mentioned above, are explicitly indicated by identifiable linguistic markers in the texts.

Our hypothesis is that semantic relations leave some discursive traces in textual document. We use the cognitive principle which on based upon the linguistic marks found in the organizing discursive relations within the text.

Semantic categorization: connection between named-entities

A semantic map represents the various specifications of the semantic relationship between concepts associated to a "point of view". For instance we present the semantic map of the "point of view" of CONNECTION (who is *connected* with who ?)

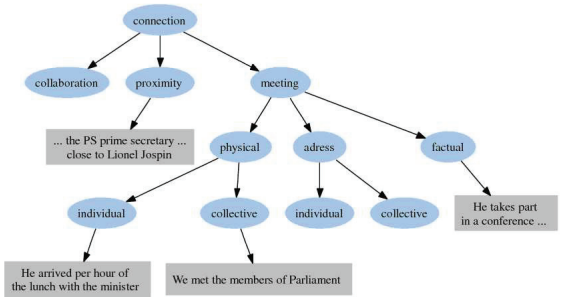

Figure 2: Semantic map for connection semantic category

The first level of the semantic map makes it possible to release three types of meetings between agents: (i) collaboration, (ii) proximity and (iii) general meeting. Connection rules are triggered by occurrence of nouns connected to a meeting verb, and the semantic annotation is assigned if linguistic clues, like spatial prepositions, are founded in the indicator's context. For example, a connection is detected in titles which contains the linguistic form composed by an indicator defined by *"a visit, a meeting, an appointment, a seminary, etc."* and the preposition clue *'s, with, ...* Sentences on title structure annotated as a factual connection by this rule are as :

(4) : Briefing on President Bush and Prime Minister Blair's ***Meeting***

In addition, the process annotation must distinguish between a generic annotation and a specific annotation. In the specific annotations, linguistic rules use ENAMEX (proper nouns and named-entities or locations, like *Prime Minister*, *Downing Street*, etc.) and TIMEX (temporal expressions). In generic annotation, rules are declared according to Contextual Exploration for an indicator (generally a verb) on a textual segment within clues expressing a connection relationship (Desclés, 2006).

An other rule for identifying a physical meeting is defined on verbs like *interviewed by, came with, ...* and a clue represented with a named-entity (MUC's ENAMEX and TIMEX). Sentences on body content structure annotated by this rule are as :

(5) : Le Premier Ministre, ***accompagné*** *du Ministre de lIntérieur,...*

(5') : President Bush ***was interviewed*** *for more than an hour yesterday ...*

Annotation engine EXCOM

A major objective for EXCOM (Djioua et ali., 2006) system is to explore the semantics and the discourse organization of texts, in order to enhance information extraction and retrieval through automatic annotation of semantic relations. Most linguistic-oriented annotation systems are based on morphological analysis, part-of-speech tagging, chunking, and dependency structure analysis. The methodology used by EXCOM, called Contextual Exploration, describes the discursive organization of texts exclusively using linguistic knowledge present in the textual context. Linguistic knowledge is structured in form of lists of linguistic marks and declarative rules for the Contextual Exploration from each linguistic mark. The constitution of this linguistic knowledge is independent of a particular domain. Domain knowledge describes the concepts and their subconcepts of a subject domain with their relationships. The contextual knowledge concerns communicative knowledge as a discursive organization, which deals with the preferences and needs of those who use information in the texts.

Linguistic rules for identifying and semantically annotating segments use different strategies defined through rules. Some of these rules use lists of simple patterns coded as regular expressions, others need to identify structures like titles, sections, paragraphs and sentences for extraction purposes. The most relevant rules for EXCOM are those called Contextual Exploration (CE) rules. A CE rule is a complex algorithm based on a prime textual mark (called indicator), and secondary contextual clues intended to confirm or invalidate the semantic value carried by the indicator.

The core of EXCOM annotation model is divided on the following interlinked parts:

(i) Textual document

(ii) General metadata like (title, author, edition, etc.)

(iii) Semantic annotations in relation with semantic categories for discourse

We realize an annotation engine for French language and extended to other languages such as Arabic (Alrahabi, 2006).

Semantic annotation processing

The first step in building a linguistic categorization is to establish backbone lists of semantic marks and contextual rules which express this notion. The major subdivisions within a semantic categorization include the structural segments: linguistic marks, search space, indicator, linguistic complement marks, and annotation specifications from a semantic map. The process of EXCOM annotation consists of the following steps (depicted in Figure Fig.3):

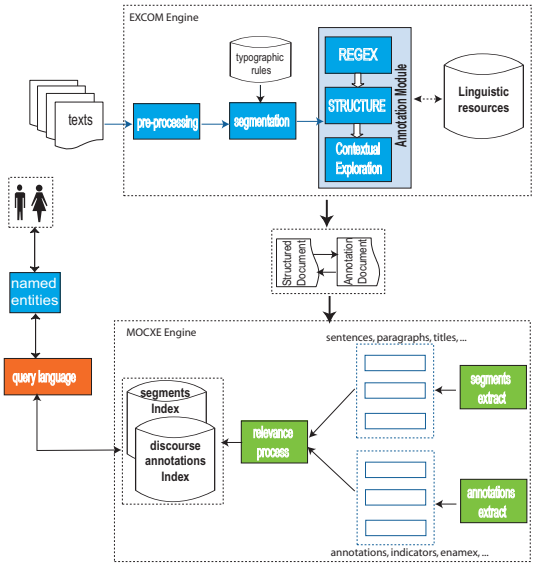

Figure 3: EXCOM/MOCXE architecture

Semantic resources

Semantic resources are composed of marker lists and typed semantic rules. Lists of linguistic terms are UTF-8 coded in simple plain text files. Linguist designers for semantic resources are allowed to use "regular expressions" as terms. Some of these lists are in some "Contextual Exploration" rules considered as indicators and represents complement marks (linguistic clues) for others. Semantic rules using the "Contextual Exploration" method express the discursive organizations for the semantics of the text. Each semantic view of the text calls up a set of semantic rules which are associated with lists of terms. They are defined in an XML file.

Each "point of view" requires semantic resources (marker lists and rules). We have presented several semantic resources for French such as (Djioua, 2006) and (Le Priol, 2006), and for Arabic (Alrahabi, 2006).

Description of the annotation engine

The core of the engine is organized on several modules interconnected for generating semantic annotations. As we see in this diagram, the dependence of the forms of the rules goes from the core towards the exterior. Rules of low level are encoded in "regular expressions" and are completely independent, so they can be alone by the annotation engine. The contextual rules, therefore high level, are based on forms of rules moreover low levels like the rules using the annotations, the lists of markers quantified and the regular expressions. Let us view the detail of these levels:

REGEX: The first module is the basic annotation with regular patterns expressed by a language of regular expression. This level is used for named-entities identification, complex structures, and some sub-segments like indicators and complements linguistic marks. Each regular expression uses one of these three levels basic, extended and advanced regex. The advanced regex level introduces Unicode with look-ahead and look-behind patterns. We add the possibility of using lists of markers and algebraic operators like *, + and ? on regular expressions. EXCOM uses the "regular expression" engine of the Perl programming language. An example of annotation rule for named entities.

This above rule represents an annotation of two contiguous words which starts with a capital letter as a proper name (Jacques Chirac, Tony Blair, etc . . .).

STRUCTURE: This level makes it possible to use pre-annotated segments as indicators or clues. This feature forces the engine to reach every annotated segment in the document structure (done with XPath expressions).

CONTEXTUAL EXPLORATION Our approach is based on the Contextual Exploration Method (Desclés et al, 1991 and Desclés 2006), which states that semantic information associated to textual segments can be identified by linguistic primary marks (called indicators) and a set of clues that would help to tackle their polysemy and indetermination.

This is the most important layer of the annotation engine. A CE triggers complex mechanisms that need the use of XSLT transformation language and a programming language (in this case, Perl). To continue with *'Prime Minister Blair'* in point of view of "connection", if a user wants to annotate a sentence like

"The British Prime Minister Tony Blair's visit to Paris last week ..."

A semantic rule based on Contextual Exploration method would follow these steps:

(i) Express the semantic of the "connection" category by means of a relevant indicator, represented in this sentence by the verb *'visit'*

(ii) To confirm the indicator's "connection semantic", we need first to identify in the text (in a sentence) the spatial expression *'at Paris'* in this right context

(iii) Indicator needs another expression like the named entity *'The british Prime Minister Tony Blair'* to allow the engine to annotate the sentence.

EXCOM uses an XSLT engine (with XPath parser) to identify nodes in the input XML document and process transformations by adding XML elements and attributes.

The annotation engine process, in our example of "connection point of view" for French language, is as follows:

(i) Identification of the indicator in the text - in a title structure for this rule - (terms of the list "NomsDeRencontre" - a list of names like *'une viste/a visit'*) - this step generates an annotated and structured text with a markup `"<indicator rules='RRencontre101a'>Visite </indicator>"`. This process also generates an XML document which represents the candidates segments for this rule.

(ii) Generation of search spaces: Parts of text (in the same title structure) where the engine will search the

linguistic clues that will confirm or invalidate the indicator's connection value (one being a list form, the other two being the pre annotated segments named entities <nom_propre> and spatial expression <expression_spatiale>). These linguistic clues are identified sequentially (ordre_entre_indices="suite"). Only the right location is generated for this rule.

(iii) The term "de/of" is the first term from the list "IndiceDetRencontre" - which is the search mechanism activated within this identifier.

(iv) Identification of the second and third clues declared in the rule as a pre annotation of named entities (<nom_propre> and <expression_spatiale>). This operation is realized with an XML tree transformation using XPath/XSLT engine. A XSLT stylesheet is applied on the previous pre annotated XML document. This process produces two outputs: the structured document and its associated semantic metadata file.

(v) Annotation generation and relationship with the segment tile.

This annotation expresses that a phrase (the title of the news paper article) is marked as a connection relationship between named-entities, with one of the agents identified as *"M. Hollande"*. EXCOM results are prepared with these two structures to be easily manipulated by final users towards graphic viewers. Programs can also use these two interconnected (XLinked structures) documents for information extraction and indexing.

Indexing annotated segments [1]

The majority of existing automatic indexing methods select index terms from the document text. The index terms selected concern single words and multi-word phrases extracted from titles and full-text of documents and are assumed to reflect the content of the text. Currently, the search engines that operate on the Internet index the documents based upon this principle. However, not all words in a text are good index terms and words that are good index terms do not contribute equally in defining the content of a text. A number of techniques help in identifying and weighting reliable content terms.

The strategy used here is divided into two parts: A first step which deals with annotation of the textual documents according to discursive point of view and finally the storage of these annotated documents; in the second step, indexing annotated segments (sentences and paragraphs), in order to provide an answer not only with a list of documents, but also with the annotated segments which correspond to a semantic based query.

The aim of this operation is to build up a multiple index composed of textual segments (sentences, paragraphs, section titles, ...), semantic discursive annotations (relations between concepts, causality, definition,...) and, if possible, named-entities (enamex, locations and timex).

Let us explain the principle through a query situation as expressed in (1). To answer the question, the search engine can not just use linguistic terms such as *"Chirac"* and the morphologic forms of the verb *"to meet"*. The notion of meeting is defined by linguistic marks (indicator and clues) organized by writing traces of the author in the text. So instead of organizing the search engine index by terms, MOCXE use a double structure composed by textual segments in a side and annotations, terms and semantic marks in another side. The MOCXE index organization is as shown below :

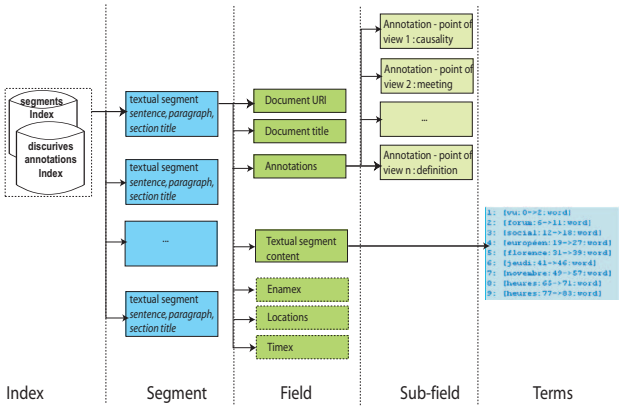

Figure 4: Organization of the MOCXE index

This organization explain the relationship between the initial document, constitutive textual segments, discursive annotations, named-entities (enamex, location, timex) and the terms which compose textual segments (sentence, paragraph and section title). A document is seen, in this organization, as a set of textual segments identified automatically by EXCOM as discursive organization of the text written by an author. Each textual segment is associated with several important pieces of information such as :

(i) a set of semantic annotation (discursive mark like *factual meeting*, *definition* or *quotation*). Each textual segment identifying as relevant for a document can be associated to a set of discursive annotations (defined by index subfields) according to the semantic point of views used in the annotation engine EXCOM. So that a same textual segment can be chosen by the search engine MOCXE as response for a query about *causality* and *quotation*.

(ii) document URI for a unique identification on the Internet.

(ii) document title for a pleasant answer.

(iii) the full-text content for a relevant answer to users. The indexing process used for a textual segment is the same usually used according to the methodology shown used by Salton.

(iv) if it was possible, enamex named-entities like proper names, organizations and trades to be able to answer *"Who met Chirac ?"*

(v) if it was possible, location named-entities to be able to answer *"Where Chirac met Poutine?"*

(vi) if it was possible, timex named-entities to be able to answer *"When Chirac met Poutine?"*

[1]The authors would like to thank B. Sauzay for his comments and who initially had proposed to index documents with Lucene.

MOCXE uses a query language which is based at the same time on both linguistic terms (constitutive of textual segments) and discursive semantic categories (definition, connection between concepts, quotation, ...) Let us see some queries for the "connection" category.

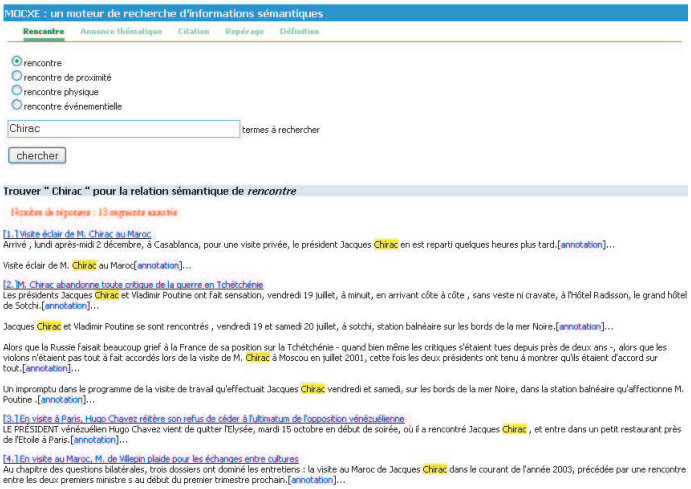

Figure 5: MOCXE answers for the query (1') in French

The answer to the query (*(1) : "Who met Chirac ?"*, in French *(1') "Qui a rencontré Chirac"*) gives a set of textual segments (body sentences and sections titles for this example) grouped through a document URI (the annotated document by EXCOM). Each textual segment presents a discursive annotation ("meeting" annotation for this example). The general process for the search engine proceed is as follows :

(i) The query, in French, has two important functions: a discursive semantic category ("rencontre/meeting") and a simple term ("Chirac"),

(ii) MOCXE extracts all textual segments founded in the index associated with the annotation *"rencontre/meeting"* (relationship segment-field-subfield in figure 8),

(iii) Selection from these textual segments, all segments within an occurrence for the term *"Chirac"* (relationship field-terms in figure 4)

(iv) Display all present information in the index related to each textual segment selected (relationship segment-field-subfield-terms and segment-field-terms).

General conclusions and future work

The notion of relevance plays an extremely important role in Information Retrieval. The concept of relevance, while seemingly intuitive, is nevertheless quite hard to define, and even harder to model in a formal way when we deal with a semantic content indexation process. We will not yet attempt to bring forth a new definition of relevance, nor will we provide arguments as to why a new definition might be theoretically superior or more complete than what has been proposed previously. We want to propose, in future work, a new formalism for modelling relevance of a document based upon semantic content for queries from the linguistic knowledge of the semantic maps associated to the "point of views" (Le Priol, 2006) (Alrahabi, 2006). The main distinguishing qualities of our formalism:

- To handle semantic discursive marks on indexed documents,
- It is necessary to use the hierarchic structural organization of the semantic point of view and the associated semantic map.
- The quality for linguistic indicators (defined by linguistic experts) affects the general score for a textual segment,

It is always difficult to evaluate a system under development and moreover identifying and indexing semantic content for web documents. The only evaluation that we undertake is the comparison of the results of a traditional search engine and this prototype about the semantic notion of "meeting". So a dual comparison can be made between the relation *document-terms* in classical indexing process and *document-textual segment-terms* in MOCXE indexing methodology. From this we have shown that existing search engines fall short of dealing correctly with queries expressed with semantic categories, while this new web search engine MOCXE has more precise search tools.

Future evaluation challenge is advancing as we build towards a qualitative comparison between textual segment chosen by the effective realized prototype in French and a human expert annotation for the same documents.

References

Alrahabi M. et al 2006. Semantic Annotation of Reported Information in Arabic *FLAIRS 2006, Floride, 11-13 Mai*

Berners-Lee Tim, Hendler James and Lassila Ora. May 2001. *The Semantic Web: A new form of Web content that is meaningful to computers will unleach a revolution of new possibilities* Scientific Americain

H. Cunningham, D. Maynard, K.Bontcheva, V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)* Philadelphia, July 2002

Desclés J.-P. 2006. COntextual Exploration processing for Discourse automatic annotations of texts *FLAIRS 2006, Floride, 11-13 Mai*

Djioua B. et al 2006. EXCOM: an automatic annotation engine for semantic information *FLAIRS 2006, Floride, 11-13 Mai*

Handschuh S. and Staab S. 2005. *Annotation for the Semantic Web* Volume 96 Frontiers in AI and Applications

Kiryakov A. et al 2004. *Semantic Annotation, Indexing, and Retrieval* Elsevier's Journal of Web Semantics, Vol. 1, ISWC2003 special issue (2), 2004.

Le Priol F. et al 2006. Automatic Annotation of Localization and Identification Relations in Platform EXCOM *FLAIRS 2006, Floride, 11-13 Mai*