Automatic Summarization of French Scientific Articles by a Discourse Annotation Method using the EXCOM System

Blais Antoine*

Hankuk University of Foreign Studies

Blais Antoine. 2009. Automatic Summarization of French Scientific Articles by a Discourse Annotation Method using the EXCOM System. *Language and Information 13.1*, 1–20. Summarization is a complex cognitive task and its simulation is very difficult for machines. This paper presents an automatic summarization strategy that is based on a discourse categorization of the textual information. This categorization is carried out by the automatic identification of discourse markers in texts. We defend here the use of discourse methods in automatic summarization. Two evaluations of the summarization strategy are presented. The summaries produced by our strategy are evaluated with summaries produced by humans and other applications. These two evaluations display well the capacity of our application, based on EXCOM, to produce summaries comparable to the summaries of other applications. (**Hankuk University of Foreign Studies**)

Key words: automatic summarization, discourse categorization, automatic annotation, discourse level analysis

1. The Human Summarization Activity

Summarization is a particularly complex cognitive task used in comprehension and memorization. Humans have the ability to summarize a large number of objects: events, oral speeches, books, movies, paintings, etc. But what are the common treatments involved in each step of the summarization process concerning these different objects? In cognitive models of summarization, the summarization process is seen as containing three main stages that we will briefly describe here (see Endres-Niggemeyer (1998)).

1. Analysis of the source of information. In the case of a text, this corresponds to the reading and comprehension of the text.

* The work presented in this paper has received financial support from the Hankuk University of Foreign Studies. We also thank the University of Paris-Sorbonne for its collaboration.

2. The central process of the summarization activity: the condensation and abstraction of the informative content of the source.

3. Production of the summary in a form chosen by the summarizer.

2. Textual Summaries

The summary, which is the result of a complex cognitive activity, has a particular function which allows one to characterize it. Beyond the various types of summary, a summary's main purpose is to give, in a condensed form, essential information about a textual object (a book, an article, etc.) so that the reader can decide what action to take based on that information (read the book or not, etc.). There are consequently various types of summaries that answer to diverse needs, but their main function remains the same as that described above. Here are the main types of textual summary:

1. *Indicative summary*: the indicative summary provides enough information to the reader so that he can judge if he wants to consult the original text or not.

2. *Informative summary*: the informative summary provides information which gives a overwiew of the contents of a text.

3. *Author's summary*: the purpose of this summary is to allow the author to introduce his text or to give some preliminary elements before reading.

4. *Summary as review*: its function is to provide a brief overview of a document filtered through the author's perspective.

5. *School summary*: the objective of the school summary is educational, and it indicates the capacities of a subject (children and students) in the reading and the comprehension of a text.

6. *Recapitulative summary*: the recapitulative summary is generally at the end of a document and its function is to provide a retrospective overview of the main elements of the text.

3. The case of professional summarizers

Professional summurizers must be able to summarize any text, whatever the area covered, and do this efficiently and rapidly. Various experiments on professional summarizers showed that they do not process the entire text through whole reading, but rather they focus on some parts (beginning and end in general), and their reading of the text is often non-linear.

Professional summarizers have specialized linguistic skills regarding the organizational structure and the distribution of information in texts. With these skills, they can distinguish the most important information that will construct the summary. In Endres-Niggemeyer (1998), professional summarizers use a common set

of strategies for finding relevant information but their text processing is different for each professional summarizer. Each seems to proceed in a different way on the same text and they do not apply the same methods at the same points in their processing. However, the criteria for relevance of the important elements seem to be nearly the same between summarizers. We note the important works of Cremins (1996) in the field of professional summarization that took care to describe the main procedures which take place in human summarization and which any summarizer has to adopt.

In the case of scientific articles, professional summarizers primarily read the beginning and the end of the text. Indeed, they search for the most relevant elements to produce the summary such as the author's aim, the presentation of the subject or the general conclusion (but also the main hypothesis, methodological remarks, problems, etc.) which are localized in these two parts of the text. The specific location of the relevant elements in the text's organizational structure corresponds to their position in the scientific argumentation that takes place in this type of text and which is reflected in parallel at the surface level of the text (the purpose and thematic presentation are mostly at the beginning of the main argumentation, the conclusion completes it, etc.).

The study of summaries of scientific articles produced by professional summarizers showed that they are generally composed of the same discourse elements. Liddy (1991) noted that summaries from a corpus of scientific articles often include the following main components: objectives, assumptions, methods, subjects, results, conclusions, and references. To recognize a textual segment as an objective or a conclusion, summarizers use various indications such as the titles and subtitles that are present in the context of the textual segment, as well as the presence of linguistic expressions - in particular metadiscoursive expressions. The identification of these linguistic expressions allows the summarizer to find and quickly extract the relevant elements, and this identification is generally performed following the selection of a textual section (first or last paragraph) by the summarizer. These linguistic expressions are not related in their use to knowledge of a particular field such as biology, physics or economics, so they can be found in any scientific article. Also, these linguistic expressions are quite independent of the author's style. All these techniques and procedures, which localize the relevant information in a text, allow professional summarizers to quickly summarize an article without the time-consuming task of full comprehension and without specialized knowledge that would need to have been previously acquired.

We will see later in this discussion that our approach to summarization, similar to that of professional summarizers, is based on the partial analysis of a text, and on the use of surface textual elements which indicate relevant information such as the author's objectives or conclusion. Our approach is thus partially based on empirical facts and established psychological hypotheses about professional summarization.

4. Automatic Summarization

There exist two major approaches to automatic summarization that we will briefly present here (see Mani (1999), (2001a) for a more detailed description of the domain

and its history).

The first approach consists in automatic summarization by *comprehension.* This approach is influenced to a large extent by models proposed in the fields of Artificial Intelligence and Cognitive Psychology. In this framework, the automatic summarization task aims to simulate the process of human summarization. Consequently, automatic summary construction relies on a stage of total or partial comprehension of the text, i.e., an in-depth analysis of the text. The applications try to create a text representation that will be later modified in order to generate the summary. A number of methods use this approach of comprehension. For example, the Susy system (Fum, Guida, and Tasso, 1982) is inspired largely by the model of comprehension of Kintsch and Van Dijk (1978), namely by the elaboration of the propositional representation of the text. Other methods aim to build, or rather instantiate, schematic structures of the text that are defined in advance in order to make a summary (use of event scripts proposed by Schank (1977)).

The second approach is that of the summarization by *extraction* with surface analysis, inspired by the domain of Information Retrieval. This approach was first introduced in the works of Luhn (1999) and Edmundson (1969). The goal of this approach is to produce a summary quickly, by simply using surface text elements and without making any deep linguistic analysis of the text. The most relevant textual segments are localized and extracted in order to obtain a subset of text extracts that is considered as the summary. In this approach we can distinguish two major ways to proceed in working with surface text elements.

Firstly, statistical methods, parallel to the models of information retrieval, take as a relevance criterion the score of a textual segment calculated as a function by using various conditions (Hatzivassiloglou et al., 2001), (Mitra, Singhal, and Buckley, 1997), (Salton et al., 1999). Therefore, a textual segment is extracted if its score is sufficiently high compared to a specified threshold and the scores of other segments. The criteria taken into consideration for the relevance evaluation are relatively heterogeneous. Among the main criteria, we have the frequency of terms that are relevant to or representative of the text, position in the text, the presence of title terms, and also the presence of some linguistic markers. These statistical methods are characterized by the fact that they work entirely with numerical values often obtained by the use of arbitrary weights or machine learning.

Secondly, linguistic methods rely on the presence of surface linguistic markers and criteria of the discursive nature of the markers, such as position in the discourse structure, in order to determine the relevance of a sentence in the text. The objective is to identify the most important segments by using linguistic knowledge (markers, discourse structure, etc.) without using a quantitative relevance evaluation based on diverse criteria. These approaches are generally based on the idea that certain surface markers could indicate in a precise textual context the semantic value of the segment in which they appear, and in this way reveal the segment's relevance in the text.

The advantages of the method that uses surface analysis for extraction is that it does not require the use of schema or textual semantic representation and can provide a summary in a simpler way, without text generation. This method is also adequate from a cognitive point of view. In fact, professional summarizers tend to

create their summaries by a surface analysis of the texts, essentially by the operations of the copy-paste type and a few supplementary modifications. Formulations that are unique to the summarizers are rather rare in the summaries (Endres-Niggemeyer, 1998).

Nowadays, works in this domain are more and more oriented toward flexible summary production corresponding to user needs. Independently of the methods that are used, different types of summaries are proposed, such as indicative, descriptive, technical, etc. The aim is to respond to a type of information retrieval need.

Moreover, the extension of automatic summarization to new applications brings up the need of summarization of a number of documents in the same processing: multi-document summarization. The methods of summary construction are generally the same but new difficulties appear, such as for example the redundancy of information that could exist in the documents and that has to be removed from the final summary.

5. Discourse Text Annotation

We will present in this article a method for discourse annotation of sentences in a text based on the presence of surface linguistic markers. It is by this method that we can distinguish the sentences in a text that belong to the discourse categories that are most relevant to a given task, such as automatic summarization.

In our work, we have focused our attention on scientific articles that contain discourse linguistic markers (expressions) that play specific roles in texts. The discourse markers correspond to different linguistic units more complex than lexical and grammatical units. They operate on a higher level than the sentence level. Below are some examples of discourse markers (continued or discontinued) that allow one to identify the textual segments in an article that explicate the contents of the text:

- *The aim of this article is ...*
- *It is important to note that ...*
- *The approach we adopt here is ...*
- *Our hypothesis is ...*

These markers are generally domain independent, because they can occur in different types of articles: linguistic articles, medical articles... We give below two sets of examples where these markers appear in the same way in different contexts relating to different fields of knowledge :

1. Computer science
    - *The aim of this article is* to show the difficulties in protecting a network with the help of classical tools.

- *It is important to note that* current software remains ineffective during massive attacks.
- *The approach we adopt here is* to prohibit unauthorized access in private networks.
- *Our hypothesis is* that all connections to the internal network are potentially dangerous.

2. Mathematics

- *The aim of this article is* to demonstrate the validity of theory Y.
- *It is important to note that* it is not possible to reduce theory Y in the formal context C.
- *The approach we adopt here is* constructivist and we will not use indirect reasoning to prove Y.
- *Our hypothesis is* that the validity of theory Y can be justified from the non-obvious premise P.

Their presence allows the extraction of textual segments that are relevant to a particular domain. Clearly, textual segments containing these markers give relevant information on the global content of the article and can consequently be retained in constructing a summary. These linguistic markers are easily identifiable and represent the author's indications to the reader that guide them in text navigation. Such markers have the function of directing the reading of the text by signaling, for example, topic presentations, hypotheses, important points, definitions, citations, specialists' opinions, etc. These linguistic markers can be organized and exploited for the automatic annotation of textual segments. They correspond also to what in a number of works are called *cue phrases* that play a similar role (see Teufel (1998)).

In scientific articles, certain categories of information (or discourse categories) are more important than others. For example from the point of view of the construction of indicative summaries, the segments containing the general problems that are discussed and the authors' objectives tend to be more relevant than the segments containing a simple example or a citation. Therefore, their localization and extraction are of priority for this type of task. And, it is necessary to construct the set of surface discourse markers expressing a relevant category of information.

The discourse categories we use for the annotation of scientific articles are listed in Table 1 with examples.

The discourse categories described here coincide partly (and do not contradict to) the categories proposed by researchers in Psycholinguistics for the summarization of scientific articles. Liddy (1991) has observed that the summaries of a corpus of articles contain the following components: objectives, hypotheses, methods and topics.

In order to annotate sentences according to their discourse categories, we use the Contextual Exploration (CE) Method (Desclés, 1997), which consists primarily in the localization of indicators in the text that correspond to surface linguistic markers. The simple presence of an indicator in a segment is not always sufficient

for its annotation, because the discourse value of the segment can vary in different contexts. The discourse value that we want to identify by a type of indicator has to be evaluated by the application of Contextual Exploration rules in order to remove polysemy. These rules consist in identifying linguistic clues in the textual context of the indicator that allow us to determine its semantic role and annotate the textual segment. More than one linguistic clue can be associated with an indicator in order to confirm or deny a given discourse value.

Let us take two examples as an illustration of the CE method.

a. *"**We propose** in this article a detailed demonstration of the disappearance of the dinosaurs."*

b. *"To give you an idea of this approach, **I propose** that you look at the image below."*

We consider the segment ***I propose*** as an indicator for an author's presentation of a topic in scientific articles. Nevertheless, its presence is not enough to consider the sentence definitively to be a presentation of topic by the author. As shown in example *b*, this segment can also occur in a sentence that does not refer at all to what is presented in the document, i.e., a topic presentation by the author on the

Discourse category	Example
Topic Presentation	*"We will present in this article a panorama of the contemporary ideas on the notion of civilization."*
Problem	*"The problem we have to solve is the following: what is the best way of optimization of the parameters of a machine learning algorithm."*
Objective	*"The aim here is to show that Aristotle was right."*
Hypothesis	*"The general hypothesis is based on the rejection of the evolution principle."*
Technical Description	*"The concept of causality is decomposed into two relations: a temporal and a conditional relation."*
Method	*"We will use a method of behavior analysis in order to locate the different viruses."*
Evaluation	*"The evaluation shows the this new system gives better results."*
Result	*"We obtain the following values for the precision: 82% and 95%."*
Conclusion	*"We conclude that this voting system is unsuitable for use in a general election."*
Consequence	*"The synthesis of this molecule leads to a modification in the energy production of the system."*
Recapitulation	*"Globally, we could summarize all this by saying that the mathematical analysis is essential for the economics modeling."*
Recall	*"We remind the reader that this experiment has already been carried out."*
Reformulation	*"In other words, the current situation is extremely dangerous."*

[Table 1] List of discourse categories

subject of the article. In example *a*, there are two contextual clues in the proposition that confirm that the discourse role of the sentence is a topic presentation. The first one is *a detailed demonstration*, which is the result of an act of thought or speech. The second one is *in this article*, which links the topic presentation with the current document. Therefore, we can be sure that this sentence is a topic presentation by the author referring to the current document. On the other hand, in example *b*, we do not have enough clues to confirm that the sentence is a topic presentation.

The CE method is essentially a linguistic and computational method that is particularly useful in semantic analysis, where the consideration of context is necessary. From a theoretical point of view, Contextual Exploration is a process of abductive reasoning, as defined in semiotic works of C.S. Peirce (Desclés, 1997). The use of Contextual Exploration is more appropriate to the analysis of linguistic phenomena than the techniques of finite state automata and transductors or regular expressions. We give below the main features that differentiate the EC method of other techniques of textual pattern recognition (Desclés and Djioua, 2008) (Blais, 2008):

- The components of an indentified pattern in a textual segment do not all have the same importance. We distinguish in CE method the indicator of indices that subordinate it. The semantic value which is sought is mainly in the meaning of the indicator.
- After identifying an indicator, the search for clues always proceeds from the indicator either to the left or to the right. In the case of regular expressions, the identification of a textual pattern is always from left to right in a linear fashion.
- It is also possible to search negative clues, i.e. clues which must not be present in a particular textual context. In classical Kleene regular expressions, there is only the complement operator which is not equivalent to the negation operator.

6. The EXCOM system

The EXCOM system uses Contextual Exploration (CE) rules (Djioua et al., 2006), that aim to annotate the text on the discourse level by searching for indicators and clues that we have compiled and categorized. The application of these declarative rules permits the system to automatically identify textual segments (here sentences) that belong to the discourse categories presented above that are considered relevant to the summary.

The different linguistic markers (indicators or clues) that are collected and elaborated by manual linguistic analyses are stored in files either as lists of words or as regular expressions. The use of regular expressions allows us to describe general patterns associated with discourse markers in linguistic expressions. Thus, the syntactic changes of (complex and discontinuous) discourse markers do not make their identification difficult, because only general patterns will be considered. In the case of the identification of topic presentation discourse markers, we have as an

```
<!-- REvaluationAuteurDocumentencours6 : evalue par - indice auteur - ici... -->

<regle nom_regle = "REvaluationAuteurDocumentencours6" tache = "resume"  point_de_vue =
"evaluation" type = "EC">
       <conditions>
              <indicateur espace_de_recherche = "phrase" type = "annotation" valeur =
"forme-evaluation2"/>
              <indice contexte = "gauche droit" espace_de_recherche = "." type =
"annotation" valeur = "forme-indice-auteur"/>
              <indice contexte = "gauche droit" espace_de_recherche = "." type =
"liste" valeur = "Deictique"/>
       </conditions>
       <actions>
              <annotation type = "ajout_attribut" espace = "phrase" annotation =
"evaluation-auteur-documentencours"/>
       </actions>
</regle>
```

[Figure 1] CE Rule in XML format

example the following patterns (manually built): *The (aim—goal—purpose—...) of (this—these)**(word)***(article—paper—book—...)**(word)***(is—will be—...)*, where ***(word)**** represents a optional sequence of one or more words.

From a technical point of view, the CE rules are written in XML language (fig. 1) and applied automatically to texts by the EXCOM system. The CE rules are written with explicit tags that facilitate the reading of these rules, especially by linguists who are also able to grasp them easily. The writing of these rules allows the linguist to put as many positive or negative clues (i.e. to be present or not) as he wants by adding conditions between the indicator and clues (presence of clues to the left or right of the indicator), or between the clues (the first clue must precede the second)... The definition of a rule also offers the possibility of describing the action that is executed at the end: annotation of a textual segment if all clues are present with the indicator, launch of new CE rules if some clues are absent...

By using linguistic resources (discourse markers and CE rules), the EXCOM system carries out the discourse annotation of the text. The input file for this system is the segmented text file in XML format obtained by the SegaTex module (Mourad, 2001), which adds XML tags to the text in order to reveal the physical structure of the text (sentences, paragraphs, sections, titles, ...). The discourse annotations attributed to sentences are represented in the form of new meta-data added to the input XML file.

If we consider other annotation systems, they fall primarily into the field of ontology engineering. The majority of these systems locates named entities in texts (dates, monetary values, places, company or organization names...) and associates these named entities with ontological classes by associations like "is a" (see for example KIM platform, Kiryakov et al. (2004)). Some of these systems provide a

completely automatic processing while others only semi-automatic. We give below some important facts about these systems (Desclès and Minel, 2000):

- Most linguistic-oriented annotation systems are based on NLP sub-processing like morphological analysis, part-of-speech tagging, chunking... (see for example the GATE framework (Cunningham et al., 2002)). Therefore, these systems are very dependent on the quality of these sub-processing routines and their execution time.
- Most annotation systems are also based on machine learning methods. They need the construction of a pre-annotated corpus, which is a major obstacle (time and human resources required for the manual pre-annotation, heterogeneity of the corpus...).
- Current annotation systems use different semantic resources: thesaurus, lexical networks (like WordNet) or semantic networks (like UMLS). The use of separate semantic resources always seems unavoidable in the annotation process.

To make a comparison, we would now like to mention some important points that characterize the EXCOM system. This method *does not* require any (Blais and Desclés, 2006):

- preliminary morphological or syntactic analysis (the method uses only surface markers without deep processing)
- named entity recognition in different domains
- domain ontologies (the discourse markers that are used are domain-independent)
- dictionaries of synonyms (WordNet) relying on resources that are difficult to obtain and keep up to date (because of the rapid evolution in some domains)
- machine learning algorithms (linguistic resources have been initially built manually by human text analysis)

The approach we adopt with the EXCOM system is, however, compatible with named entity recognition, domain ontologies, auxiliary morphological and syntactic analyses, dictionaries of synonyms, statistical methods, etc. The aim of the EXCOM system is not to provide an alternative to other existing methods in NLP, but rather to propose a complementary approach, perhaps more adapted to respond to some specific tasks, such as information retrieval of discourse categories.

7. Presentation of the Summarization Process

Our summarization strategy is based on the method for automatic discourse annotation that we have presented above. In fact, the main criterion of relevance evaluation of the textual segments is the discourse annotation that is attributed

to the segment. According to the type of summary (indicative, informative, etc.) and the user's needs, some types of annotations are privileged more than others for the summary. Together with discourse annotation, we use two other criteria in order to evaluate the relevance of a segment. These are its position in the text and the presence of thematic terms in the sentence. These two criteria are frequently used in other approaches (Edmundson, 1969), (Teufel and Moens, 1999). Position in the text corresponds to the place of the segment in the physical structure of the text (beginning, end, etc.) and is indicative of the relevance of the segments annotated with certain categories - for example, the conclusions. In our approach the thematic terms are the terms that are most representative of the subject of the article. Presently, they correspond to the nouns found in the title that we extract automatically. Note that it is the discourse annotation attributed to the sentence that constitutes the main criterion of relevance evaluation. Position in the textual structure and the presence of thematic terms are only additional clues to the relevance evaluation of the sentence.

The summarization strategy that we adopt consists in the selection of the sentences belonging to discourse categories that conform to the expectations of a prospective reader. The construction of flexible summaries is based on the possibility of choosing the discourse categories that will construct the summary. Therefore, a number of strategies are proposed to the user (Blais, Atanassova, and Desclés, 2007), (Blais, 2008). For example, for the construction of a indicative summary of a scientific article, one of the proposed strategies is to give priority to the categories of Topic Presentation, Objective and Problem, which are situated at the beginning of the text, and the categories of Conclusions and Recapitulation at the end of the text.

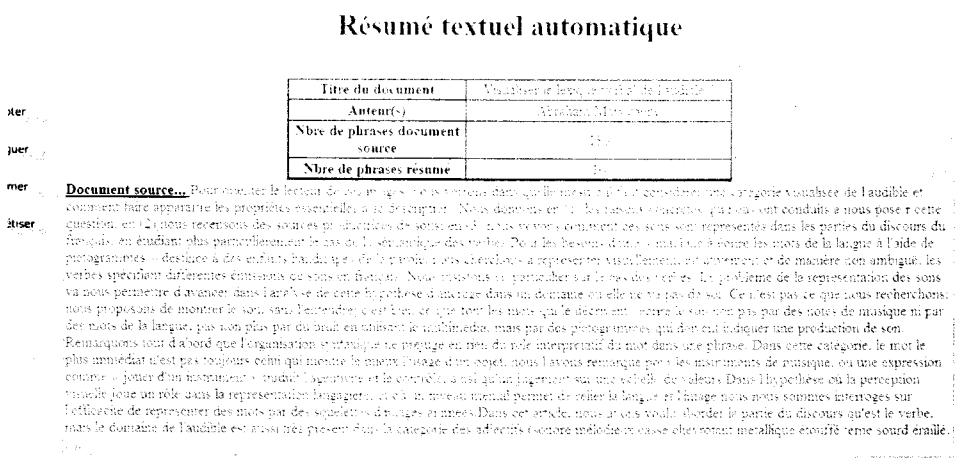
We present here the steps in the process of summary construction.

Step 1 *Thematic terms extraction.* We find and extract the thematic terms, i.e. lexical terms that are representative of the subject of the document, from the titles and sub-titles of the document, which have been automatically localized during the stage of segmentation. The terms contained in the main title of the article have the greatest importance.

Step 2 *Automatic discourse annotation of the text according to the discourse categories relevant for summarization.* The EXCOM system detects different linguistic markers in the text and applies Contextual Exploration rules in order to annotate the textual segments.

Step 3 *Initial cleaning-up of the discourse categories that are not relevant for the summary.* We remove the segments that are annotated with categories that we consider as not important: examples, citations, etc.

Step 4 *Summarization strategy.* We evaluate, depending on the summary type, the relevance of each of the annotated sentences in the text according to their discourse categories, their position in the textual structure, and the presence of thematic terms. As mentioned above, some discourse categories are more relevant than others; we rank-order them according to a prede-



[Figure 2] Visualization of a summary

fined hierarchy for each particular strategy. This hierarchy varies according to the type of summary that the user wants to obtain. For example, in the case of indicative summary, Topic Presentation is at the top of the hierarchy.

Step 5 *Cleaning-up of the summary.* Finally, we proceed to clean-up the summary in order to improve its cohesion and readability. For example, we remove the enumerations (*firstly, secondly...*), or some connectors in the beginning of the sentence (*Then, And...*), as they could reduce the readability of the text by showing inconsistencies in the summary's cohesion.

Step 6 *Visualization.* The user can visualize the summary in a dedicated interface (fig. 2) where the various discourse categories of the summary can be shown. It is also possible to display some additional information related to each sentence: description of the discourse category, position in the textual structure, thematic terms, etc.

8. Evaluations

For the two evaluations of our approach to automatic summarization, we use an intrinsic quantitative method (see Mani (2001b) for a definition of the intrinsic and extrinsic evaluation methods).

In the first evaluation, our aim is to compare the summaries produced automatically by our system and other applications to human summaries of the same texts produced by extraction. The evaluation is made on texts in French, chosen randomly out of a corpus of articles from the following French conferences and

Text	Sentences	Words	Corpus	Domain
1	*310*	*6091*	*ALSIC*	*Education*
2	*306*	*7875*	*CALS*	*Natural Language Processing*
3	*238*	*5207*	*AFIA*	*Artificial Intelligence*
4	*193*	*3690*	*TALN*	*Natural Language Processing*
5	*222*	*5232*	*AFIA*	*Artificial Intelligence*
6	*222*	*4531*	*TALN*	*Natural Language Processing*
Total	*1491*	*32626*	-	-

[Table 2] List of random selected texts

reviews: AFIA, RECITAL, CALS, TALN and ALSIC (see Blais (2008) for full details). Among the subjects of these texts are Artificial Intelligence, Linguistics, NLP and Science of Education. We want to show that our approach, which uses discourse annotation and the selection of annotated segments following a given strategy, is efficient for these types of texts, especially in comparison with other software applications for automatic summarization that use different methods.

The reference summaries were constructed by extraction by Master students at the university of Paris-Sorbonne. Each student was presented with a text and asked to select a set of sentences, representing 10% of the text, that are the most relevant for the construction of a classical indicative summary. The texts were around ten pages long and the students had one hour to complete the task.

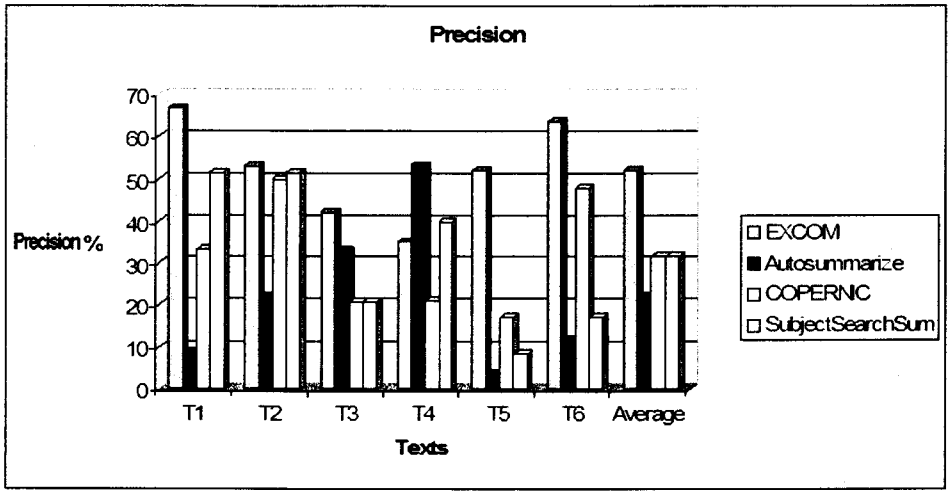
Twelve students participated in this evaluation, and six texts were summarized in total (Table 2). Each text was summarized by two students, which allowed us to calculate the coverage rate[1] between the two extracts. In fact, for a comparative evaluation, working on a single reference summary would not be justified, because of the considerable differences between human summaries of the same text. The average coverage rate we obtain for the six texts is 33.12%. That means that generally, two students working on the same text agree on the extraction of one sentence out of three.

Independently of the students, we have produced automatic summaries of the same texts, by using our method presented above and three other software applications for automatic summarization. These three applications are professional applications that offer the functionalities of automatic summarization of texts in French:

1. AutoSummarizer by Microsoft Corporation (2005 version), provided in Microsoft Word
2. Copernic Summarizer by Copernic, Inc. (2005 version)
3. SubjectSearchSummarizer by Kryloff Technologies, Inc. (2007 version)

These three applications, like almost all others, mainly use frequency criteria to identify relevant segments, and some additional heuristics, such as localization

[1] The coverage rate represents the percentage of sentences extracted by both students working of the same text.



[Figure 3] Values of the precision of automatic summaries

in the textual structure or the identification of relevant expressions built on a finite corpus by machine learning. However, none of these applications seems to rely on linguistic resources (rules, markers...) constructed manually and therefore none of them proposes discursive categorization of textual segments.

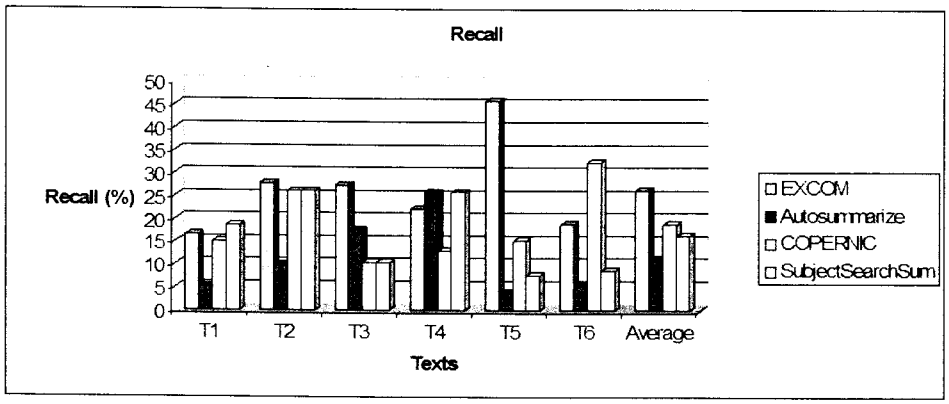
We have quantified the values of precision for each of the summaries produced automatically, by using the human summaries as reference. The precision here corresponds to the number of sentences in the automatic summary that have been selected by the student divided by the total number of sentences in the summary. We present in Fig. 3 precision values of for the six texts and each of the applications.

We have also quantified the values of recall for each automatic summary. At first, we used as reference the union of the two sets of sentences extracted by the students for each text. Secondly, we used the intersection of the two sets of sentences, i.e., the sentences that have been extracted by both students working on the same text. We consider the second set of values we obtain the recall to be more revealing of the capacities of the automatic summarization systems. The results for the recall values are presented in Figs. 4 and 5.

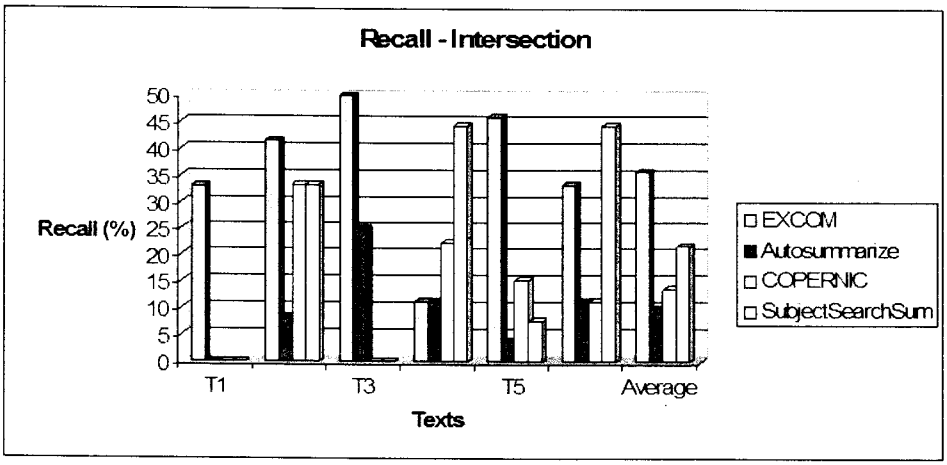
In comparison to other automatic summarisation systems, our approach has a leading position according to the three values that have been calculated for scientific articles.

We can make some important remarks on the obtained results:

- Considering precision, the summaries produced by our system contain an average of 52% of the sentences chosen by the subjects (values of precision in Fig. 3), this value being much higher than the values for the other three systems. It must be noted however, that in this case the precision value does not necessarily indicate the quantity of noise, as the automatically



[Figure 4] Values of the recall on the union



[Figure 5] Values of the recall on the intersection

extracted sentences that are not extracted by the subjects are not always noise.

- An analysis of the noise revealed that the three other systems have included in the summaries sentences that are totally irrelevant from the point of view of a reader, such as a random section title of the article or a simple bibliographic citation. This is probably due to the use of statistical methods and heuristics, which in some cases give this particular kind of noise (because there are no justifiable linguistic criteria for the selection of such sentences). In our approach, the noise instead corresponds to sentences whose discourse value has not been correctly determined by the rules because of polysemous markers. However, these sentences affect the readability of the summary less than in the case of the noise elements extracted by the other three systems.

- As for the two evaluations of recall, especially in the second case, our approach is also situated in the first position. For all the six texts, it has extracted the sentences extracted by the two subjects with an average recall of 35% (fig. 5). The other systems in some cases have not at all introduced such sentences in the summary (for texts T1 and T3). In our evaluation, the sentences extracted by the two subjects at the same time often contain discourse markers (such as meta-discourse elements) that explicitely indicate their relevance. Generally, the segments selected by human summarizers are often chosen because of their discursive category (Liddy, 1991). Summarizers select segments belonging to the most relevant categories and their identification is possible from linguistic markers (discourse markers) (Cremins (1996), Endres-Niggemeyer (1998)). It is for this reason that statistical methods employed by the majority of applications in automatic summarization are inefficient to identify segments selected based on these linguistic criteria. The low results of recall values obtained for the three profesionnal applications show the limitations of using of statistical methods based on high-frequency words and also justify the use of linguistic techniques to identify relevant segments like human summarizers do.

- Finally, we could see in the three figures that generally, the variations in the values from one text to another are smaller for our approach than for the others. This means that our strategy is more reliable for the processing of different texts, unlike AutoSummarizer, for which these variations are rather large, showing its unreliability.

In the second evaluation, five summaries of the same text (with a compression ratio of 10 %) were given to twelve other subjects (Master's students at the University of Paris-Sorbonne). We divided the group of subjects into two subgroups, where each subgroup processed five summaries of a different text. For each summary, a student had to attribute a qualitative grade of either *poor*, *medium* or *good* (giving the quantitative values 0,1 or 2) based on specific criteria (indicative function, coherence, cohesion, etc.). The students had no access to the original documents and

	Copernic	MsWord	Excom	Human 1-2
Total average	1.00	0.75	1.00	1.13
Average deviation	0.52	0.66	0.40	0.51

[Table 3] Average values for the two texts

they were not considered to be experts in the fields covered in the texts. In the five summaries given to each student, two had been composed by human subjects, and the three others were produced by a software application. Both initial texts of this evaluation were chosen randomly from among the six texts of the previous evaluation, so the human summaries used here came directly from the previous evaluation. The three non-human summaries were generated automatically as follows: the first one by our application, the second by Copernic Summarizer, and the third by AutoSummarizer. Copernic Summarizer and our application had shown the best results in the previous evaluation and their final results were close, which is why we selected Copernic Summarizer in particular, in order to see if a difference appeared more clearly in this new evaluation. To avoid tiring the subjects out during the evaluation process, we did not select more software generated summaries. Students had one hour to evaluate five summaries, since we believed that the reading of more summaries could have impaired the quality of the evaluation. Finally, to avoid a possible statistical bias due to the order of presentation of the five summaries, and assuming that the evaluation for the later summaries would be different from the evaluation of the first ones due to change in reader's focus, we presented the five summaries in a different order to each student.

In Table 3, we have given the average notes and the average deviation for the two texts and for each application, and in the last column, we have the average values for the set of human summaries (Human1 + Human2 for the two texts). The results are calculated from the numerical values 0, 1, or 2 (*poor*, *medium* or *good*) assigned to each abstract by a subject. Finally, we refer to Blais (2008) works for the full details of the data obtained for each summary in this evaluation.

Let us make some remarks on the results:

- On average, it is seen in the two texts that one of the two human summaries always received the best grade, with the two automatic summaries (ours and that of Copernic Summarizer) being placed before the other human summary. This shows that in the evaluation by giving grades to the summaries, the human summaries are not so easilly distinguished compared to the automatic summaries.

- Only 2 out of 10 summaries have been graded in the same way by 6 students. The other 8 summaries, both the human and automatic summaries, have obtained grades varying from *poor* to *good* (from 0 to 2). Thus, there is no clear distinction between the human and the automatic summaries.

- If we consider the values of the average deviation, we see that the students

do not agree more on the grading of a human summary than the grading of an automatic summary.

- The summaries produced by our application and by Copernic Summarizer are not really distinguishable by the students from the human summaries (if we take the average values for Human 1-2 on table 3). The advantage goes rather to our application because of the rather small variation between grades (deviation of 0.4) for an average grade of 1.

We note that it is difficult to generalize these results because of the limited number of texts and subjects that have participated in this evaluation. However, as this evaluation was made after the first one, it allowed us to continue to observe certain tendencies that have been attested in the first evaluation. Firstly, our application confirms its capacities to produce summaries comparable to those of other applications using other methods (statistical, etc.). Secondly, our application placed slightly ahead of the others, especially compared to Copernic Summarizer, which has a higher deviation (therefore the summaries it produces are less reliable in quality). Finally, compared to the other two automatic summaries, AutoSummarizer continues to show large variance in the results, which confirms the first results showing its unreliability.

9. Conclusion

While the first evaluation allowed us to evaluate the content of the summaries, in the second evaluation, the students had to analyse the coherence and cohesion of each summary rather than the relevance of the information that they contained. During the first evaluation, our system produced summaries with a precision and recall much higher in some cases than Copernic Summarizer and AutoSummarize. These large variations did not appear distinctly in the second evaluation. The students judged the coherence and cohesion of the summaries by normal reading, without taking into consideration the relevance of information that could permit us to better distinguish between the summaries. Apparently, the students could not easily distinguish the quality of the summaries, which is revealed by the smaller devations in the results. Moreover, in most cases the students did not graded the summaries in the same way. This shows that the students had to evaluate aspects (coherence, cohesion, etc.) that are difficult to evaluate for this kind of summaries. This is evidence of the importance for the reader of the textuality of the summaries produced automatically. The fact that the summaries contain relevant information (according to the first evaluation) does not necessarily guarantee good results in the second evaluation and vice versa. For example in several cases the subjects gave an equal grade (or even higher) to a summary produced by AutoSummarizer compared to a human summary (which serves as a benchmark for the first evaluation, where AutoSummarizer obtained bad results).

Finally, if the better position of our application is more prominent in the first evaluation, the second one does not deny it, but shows smaller variations in the results that make it more difficult to rank the different automatic summaries. These two evaluations display well the capacity of our application, based on EXCOM, to

produce summaries comparable to the summaries of other applications. Generally, our application has even proved to be better in certain aspects (for example, the content is more similar to that of the human summaries). Thus, the different hypotheses and the methods that we have used for summary construction in our application (discourse annotation, Contextual Exploration method, etc.) have proved to be efficient and operational for such a difficult task.

Our objective is to upgrade our software application based on the latest developments of the EXCOM system (http://www.excom.fr), and also to apply this work to the Korean language as we did with French. Currently, our work on the Korean language focuses on the development of linguistic resources needed to build summaries in Korean, as well as validating them.

<References>

Blais, A. 2008. *Résumé automatique de textes scientifiques et construction de fiches de synthése catégorisées : approche linguistique par annotations sémantiques et réalisation informatique.* Ph.D. thesis, Paris-Sorbonne University.

Blais, A., I. Atanassova, and J-P. Desclés. 2007. Discourse Automatic Annotation of Texts: an Application to Summarization. In *The 20th International FLAIRS Conference*, Florida.

Blais, A. and J-P. Desclés. 2006. L'annotation discursive de textes et une application résumé automatique. In *Colloque : Que faisons-nous du texte ?*, Paris.

Cremmins, E. T. 1996. *The Art of Abstracting.* Information Resources Press, Arlington.

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

Desclés, J-P. 1997. Systèmes d'exploration contextuelle. *Co-texte et calcul du sens.*

Desclés, J-P. and B. Djioua. 2008. La recherche d'informations par acccés aux contenus sémantiques : vers une nouvelle classe de Systémes de Recherches d'Informations. *Linguistique Informatique.*

Desclès, J-P. and J-L. Minel. 2000. Résumé automatique et filtrage sémantique de textes. *Ingénierie des langues.*

Djioua, B., J. Garcia-Flores, A. Blais, J-P. Desclés, G. Guibert, A. Jackiewicz, F. Le Priol, L. Nait-Baha, and B. Sauzay. 2006. EXCOM: an automatic annotation engine for semantic information. In *The 19th International FLAIRS Conference*, pp. 285–290, Florida.

Edmundson, H. P. 1969. New methods in automatic extracting. *Journal of the Association for Computing Machinery* 16, 264–285.

Endres-Niggemeyer, B. 1998. *Summarizing Information.* Springer, Berlin.

Fum, D., G. Guida, and C. Tasso. 1982. Forward and backward reasoning in automatic abstracting. In *Proceedings of the 9th conference on Computational linguistics*, pp. 83–88.

Hatzivassiloglou, V., J. L. Klavans, M. L. Holcombe, R. Barzilay, M-Y. Kan, and K. R. McKeown. 2001. SIMFINDER: A Flexible Clustering Tool for Summarization. In *In Proceedings of the NAACL Workshop on Automatic Summarization*, pp. 41–49.

Kintsch, W. and T.A. Van Dijk. 1978. Toward a Model of Text Comprehension and Production. *Psychological Review* 85, 363–94.

Kiryakov, A., B. Popov, I. Terziev, D. Manov, and D. Ognyanoff. 2004. Semantic annotation, indexing, and retrieval. *Journal of Web Semantics*, pp. 49–79.

Liddy, E.D.R. 1991. The Discourse-Level Structure of Empirical Abstracts: An Exploratory Study. *Information Processing and Management* 27, 55–81.

Luhn, H.P. 1999. The Automatic Creation of Literature Abstracts. *Advances in Automatic Text Summarization.*

Mani, I. 2001a. *Automatic Summarization.* John Benjamins Publishing Company, Amsterdam.

Mani, I. 2001b. Summarization Evaluation: An Overview. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization.*

Mani, I. and M.T. Maybury. 1999. *Advances in Automatic Text Summarization.* MIT Press, Cambridge.

Mitra, M., A. Singhal, and C. Buckley. 1997. Automatic Text Summarization by Paragraph Extraction. *Compare.*

Mourad, G. 2001. *Analyse informatique de signes typographiques pour la segmentation de textes et l'extraction automatique des citations.* Ph.D. thesis, Paris-Sorbonne University.

Salton, G., A. Singhal, M. Mitra, and C. Buckley. 1999. Automatic Text Structuring and Summarization. *Advances in Automatic Text Summarization.*

Schank, R. C. and R. P. Abelson. 1977. *Scripts, Plans, Goals and Understanding.* Lawrence Erlbaum Associates.

Teufel, S. 1998. Meta-discourse markers and problem-structuring in scientific articles. In M. Stede, L. Wanner, and E. Hovy (eds.), *Proceedings of the ACL-98 Workshop on Discourse Structure and Discourse Markers*, pp. 43–49.

Teufel, S. and M. Moens. 1999. Argumentative classification of extracted sentences as a first step towards flexible abstracting. *Advances in Automatic Text Summarization.*

Submitted on: April 2, 2009
Accepted on: June 9, 2009