Contextual Exploration Processing for Discourse Automatic Annotations of Texts

Jean-Pierre Desclés
Université Paris-Sorbonne, LaLICC (Paris-Sorbonne - CNRS)
28, rue Serpente, 75006, Paris, France
Jean-pierre.Descles@paris4.sorbonne.fr

Abstract*
This paper presents a new method of annotation called Contextual Exploration which takes into account context. This method has been developed in LaLICC laboratory at the Sorbonne, France.

1. Introduction

Automatic annotation of a text is a computational process which gives a label to a linguistic unit (word, sentence, paragraph, title of a text…). The label can be morphemic (category of a word), a syntactical categorization, a semantic classification (for instance a semantic classification of verbs) or a discursive information (for instance an indication about an organization of discourse).
For more than ten years (Desclés 1991, Desclés 1997, Desclés 2001), we have defined and developed a new linguistic technique for taking into account context. This technique of Exploration of Context (EC) has been thought as a set of declaratives defined by Artificial Intelligence. Thus, it is very easy to express and to implement these rules in a computational device. Exploration of Contexts rule (EC) is presented with this general form: "IF Conditions THEN Action". The result of an action, in an annotation processing, is a specific label given to a specific linguistic unit (for instance a sentence) or the call of a new rule. In a EC rule, the condition is a set of different linguistic indices; when these linguistic indices have been found in a specific research space (for instance a sentence) then the rule can be applied.

2. Hypotheses for Contextual Exploration and Experimental Facts

The approach by Contextual Exploration assumes different hypotheses and facts.
Cognitive Hypothesis 1 : In an information retrieval processing, the looking at a text focuses only on some textual units and organizations from a discursive viewpoint.

Linguistic Hypothesis 2 : Some linguistic units are the markers of a specific viewpoint.

Fact 1 : The linguistic markers are often polysemic units.

Linguistic Hypothesis 3 : There are contextual linguistic indices which allow to remove the imprecision of polysemic units.

Fact 2 : The identifications by linguistic patterns of finite state automata are not sufficient.

Fact 3 : The linguistic identification of linguistic indicators of viewpoints are not sufficient : they add noise.

Linguistic hypothesis 4 : By an examination of the linguistic context, an automatic information retrieval system is able to find linguistic indices for reducing noise.

Linguistic Hypothesis 5 (*hierarchical principle)* : Among indices in the conditions EC rule, there is a hierarchy between one principal, called "Indicator", and other indices, called "complementary indices". The different indices (Indicators and complementary indices) are related to a specific viewpoint for an automatic annotation.
The general form of EC rule becomes :

> IF an Indicator IND, relative to a viewed point is found, THEN

IF specific indices I1, I2, …,In, relative to the same viewpoint are found in the context of the occurrence of the Indicator IND THEN
an annotation of a linguistic unit is realized ELSE
a new exploration rule is called.

An EC systems works as follows :

1°) when the Indicator IND is identified in a text then

2°) the occurrence of the Indicator IND calls all rules whose the same Indicator IND belongs to the conditions of these rules;

3°) when all complementary linguistic indices I1, I2, …, In, of a rule are found in the context of the occurrence of IND, then the rule is applied and can be conclude to an annotation of a linguistic unit.

For the identifications of associated indices determined by an Indicator in a rule, we must specify : (i) contextual research spaces where complementary indices of an Indicator can potentially found and (ii) the textual type of the labelled linguistic unit (paragraph, proposition, title, sentence …) by the EC rule.

3. An Example : Topic Announcement for Summarizing

Let us give an illustration of the general form of a Contextual Exploration rule by a figure : when an occurrence of a linguistic Indicator is found, then in a specific research space, the rule finds different linguistic indices, I1, I2 … at left of the occurrence IND, I3, I4… at right of the same occurrence and the annotation process is realized (see figure 1).

The figure 2 gives an example of a rule for an annotation of a "topic announcement" in a text.

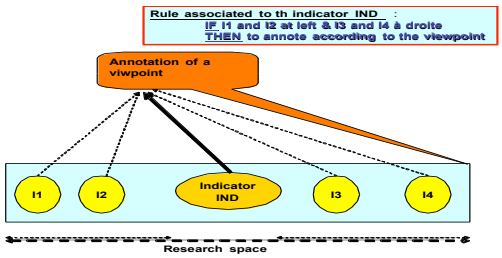

Figure 1 : Rule for annotation from an occurrence of an Indicator and different indices in a contextual space.

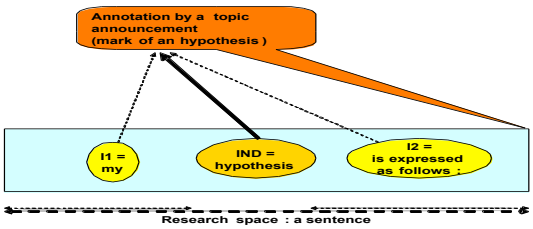

Figure 2 : Annotation by a topic announcement from the occurrence of an Indicator = "hypothesis" and different indices I1="my" at left, I2="is expressed as follows" at right.

4. Contextual Exploration Systems and Finite Automata Devices

Contextual Exploration systems are more powerful than finite automata devices.
Systems based on finite states automata are simplified examples of EC systems : in these formal approaches, Indicator IND and a string of complementary indices I1 I2, …, In are at the same level and are not integrated in an hierarchical dependence. Furthermore, complementary indices I1, …, In, in a EC rule, can be located at a very long distance from an Indicator IND (for instance in the title of an article). Thus, the expressive powers of finite states automata and EC systems are very different. Indeed, it is possible to prove that strings as « a^nb^n » and « $a^nb^nc^n$ » ($n > 1$) are recognized by EC systems with recursive EC rules but no by finite states automata.
Let us give a proof for recognizing a string as "$a^nb^nc^n$ ". Such strings are recognized by a Contextual Exploration system where the rule R1 is recursive. : when "b" is

identified as an Indicator, then the rule tries to find an occurrence of “a” in the left context and an occurrence of “c” in the right context. When we obtain a success, the occurrences change and become respectively “a’ ”, “b’ ”, “c’ ”. If there is no new occurrences of “a”, “b” and “c”, then , after deletion of mark ups, we have recognized the string given in input as a string with the same number of “a”, “b” and “c”. A such string belongs to a formal language of the family of Context Sensitive Languages (see figure 3). Such languages cannot be recognized by finite automata. Thus, the Contextual Exploration is more powerful than regular languages or equivalent devices.

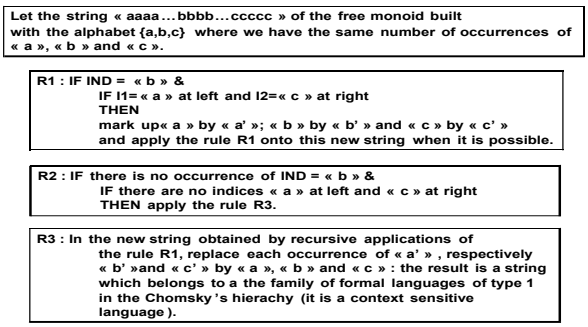

Figure 3 : Rules for a recognition of the string $a^n b^n c^n$ by a Contextual Exploration device.

5. Some Annotation Viewpoints in Contextual Exploration Approach

The building of an EC system for automatic discourse annotations in texts is related to a viewpoint yielding to annotations of this form “this sentence expresses this discourse viewpoint …”. The different previously studied viewpoints (see bibliography Desclés and alli 1991; Desclés, and alii, 1997; Bertin and alii 2006; Le Priol and alii, 2006) are :

“topic announcement”,
“conclusive remark”,
“definition”,
“report of another speaker”,
“quotation”,
“meeting between a person and other persons or organizations”,
“temporal relations between events expressed in a text”,
“causality relations between situations” …

Each viewpoint is described as follows :

1°) on the one hand, a complex relation between concepts inside a structured “semantic map” and on the other hand, a set of classes and subclasses of linguistic units (Indicators and indices);

2°) a set of EC rules where each rule relates a class of Indicators with different classes of indices.

For a specific discourse viewpoint, the associated “semantic map” is a lattice of unary concepts and binary relations, with different types (functional types), and related by arrows expressing specifications or generalizations. The classes and the subclasses are associated with every concept and relations of the “semantic map” and are structured from arrows of the map. Each class contains all linguistic expressions which are directly or indirectly associated to elements of the semantic map; these expressions (words, collocations …) are the linguistic indices which yield sufficient information for giving a label to some textual unit (a sentence or an other textual unit). Semantic map, associated classes of indices and EC rules are built together. They constitute linguistic resources for operational automatic annotations following different discourse viewpoints (see figure 4).

The semantic map is like an organization “in intension” of a discourse viewpoint, whose the classes of indices are extensional counterparts. The semantic map can be conceived also as an ontology of discourse categories, independently of different domains of application. Indeed, the expressions of the semantic map for a discourse viewpoint are the same in different domains : scientific articles or narrative texts or informative texts … since these expressions are used by the speaker to express a discourse organization of information. In some type of texts, some discursive organisation expressions will not be present but in others these expressions organize the text and give information about the intentions of the speaker.
The noise caused by the polysemy of an Indicator is filtered by the EC rules which research complementary indices in its context to eliminate false interpretations. For instance, in the mapping between a grammatical morpheme and its meaning, the context becomes necessary for building the exact meaning associated with the occurrence of the morpheme (for instance, a tense morpheme in, Desclés and alii 1991).

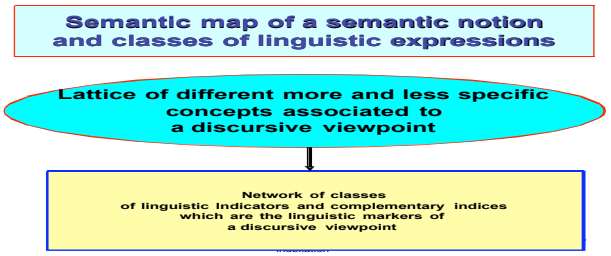

Figure 4 : The relation between a semantic map and classes of linguistic markers

The notion of EC is useful to create, by operational ways, different applications. We have used this linguistic and computational technique for the extraction of sentences through the identification of pertinent linguistic units for summarizing (for instance : *the topic of this article is* / *Principal results of my experience are …*/) and implement them in a first system SERAPHIN and in the platform ContextO, implemented with JAVA classes (Desclés and alii 1997; Minel and alii 2001; Desclés and alii 2005). Now, we are currently using this approach for annotating texts from different discourse viewpoints in the new platform EXCOM (EXploration of COntext for Multilinguist applications) with computational techniques based on XML, XSLT ... . This platform EXCOM (Djioua and alii 2006) contains different linguistic resources (classes of Indicators and complementary indices with exploration rules to relate indicators and indices) and an engine whose aim is the annotation of textual units in applying exploration rules when an Indicator of a discursive viewpoint is identified in a text. Every linguistic resource is associated with a discursive viewpoint (see figure 5).

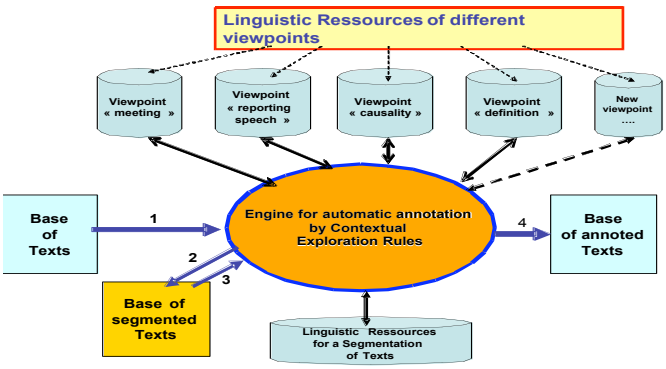

Figure 5: Functional description of the platform EXCOM for an automatic annotation of texts

References

Bertin, M., Desclés, J-P., Djioua, B., and Krishkov, I., 2006. "Automatic Annotation in Text for Biblimetrics Use", *presented at FLAIRS 2006*.

Desclés, J-P., Jouis, C., Oh, H., and. Maire-Reppert, D., 1991. "Exploration contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte ", in D. Herin-Aime and alii (eds) *Knowledge modeling and expertise transfer*, 371-400.

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

Desclés, J-P., Cartier E., Jackiewicz A. and Minel J-L., 1997. "Textual Processing and Contextual Exploration Method ", in *Context'97*, Rio de Janeiro, 189-197.

Minel, J-L., Cartier E., Crispino G., Desclés, J-P., Ben Hazez S. and Jackiewicz A., 2001. "Résumé automatique par filtrage d'informations dans les textes. Présentation de la plate-forme Filtext ", *Technique et Sciences Informatiques*, n°3, 369-396.

Desclés, J-P. and Minel, J-L., 2005, "Interpréter par exploration contextuelle ", in F. Corblin and C. Gardent, *Interpréter en contexte*, Hermès, Paris, 305-328.

Djioua, B., Garcia Flores, J., Blais, A., Desclés, J-P., Guibert, G., Jackiewicz, A., Le Priol, F., Nait-Baha, L. and Sauzay, B., 2006. "EXCOM : AN Automatic Annotation Engine for Semantic Information", *presented at FLAIRS 2006.*

Le Priol, F., Blais, A., Desclés, J-P., Djioua, B., Garcia Flores, J., Jackiewicz, A., Nait-Baha, L. and Sauzay, B., 2006. "Automatic Annotation of Localization and Identification Relations in platform EXCOM", *presented at FLAIRS 2006.*