Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts
Ágnes Sándor
Xerox Research Centre Europe

Abstract
The importance of the role of metadiscourse is increasingly recognized for natural language processing applications like text-mining and information extraction. Thus the detection of metadiscourse has recently been identified as a task in several domains, including the processing of scientific literature. We have developed a natural language processing system that detects and highlights in biomedical research abstracts a particular kind of metadiscourse that conveys the author’s rhetorical strategy. In this paper we describe the model of rhetorical metadiscourse underlying the system. Our model, combining and extending previous discourse analysis methods and models, is based on both conceptual and syntactic analyses of metadiscourse. We argue that this model is effective for automatic processing.

Résumé
L’importance du métadiscours est de plus en plus reconnue pour les applications de traitement automatique des langues comme la fouille de textes ou l’extraction d’information. La détection du métadiscours est ainsi devenue une tâche à part entière dans plusieurs domaines, notamment dans le traitement de la littérature scientifique. Nous avons développé un système de traitement automatique de langues naturelles pour détecter le métadiscours particulier qui véhicule la stratégie rhétorique de l’auteur. Dans cet article, nous décrivons le modèle de métadiscours rhétorique sous-jacent au système. Ce modèle, qui s’inspire des différentes traditions, allie une approche conceptuelle à l’analyse syntaxique. Nous cherchons à montrer que ce modèle est efficace pour le traitement automatique.

1. Introduction
A major challenge that scientific communities face today is the simultaneous proliferation of the literature and that of the channels of communication. This leads to a difficult situation for both the authors and the public: the authors have difficulties to reach their relevant readers and the readers have difficulties to find the relevant literature, as well as to cope with its abundance. This phenomenon stimulates various disciplines dealing with written communication to work out diverse strategies and methods to help researchers – both as authors and as readers – to cope with these difficulties: genre analysis is applied to elaborate adequate communication standards that help the authors effectively transmit their ideas (e.g. Kaldor et al. 1996; Swales 1990, 1994); information retrieval provides the readers with effective search methods (e.g. Google Scholar, Citeseer), information extraction (e.g. Tsujii et al. 2001; Feldman et al. 2002) or text mining (e.g. Hearst et al. 2007) propose tools that serve to accelerate the processing of large amounts of literature.

Our work takes the reader's side by tackling the issue of the actual reading of research literature once it has been retrieved. On the demand of text-mining experts in the biomedical domain, we have developed a natural language processing system that highlights metadiscourse that conveys the author’s rhetorical strategy in a collection of abstracts (see Lisacek et al. 2005). The underlying consideration that has led to the conception of such a
system has been twofold. On the one hand, it is assumed\(^1\) that drawing the readers’ attention to the rhetorical development of the text accelerates the process of comprehension, and thus the readers can more rapidly judge if a particular research article is relevant for their purposes.\(^2\) On the other hand, the content of the metadiscourse allows the reader to evaluate the relevance of the article for his specific purposes.

In this paper, we will present the theoretical bases of this natural language processing system as well as its application in biomedical abstracts published in Medline. The paper is built up as follows. In Section 2, we analyze an abstract from Medline in order to give examples of instances of rhetorical metadiscourse and show how it conveys rhetorical strategy. In Section 3, we overview the two main traditional approaches to discourse analysis, and argue that they both yield partial models of rhetorical metadiscourse. In Section 4, we propose concept-matching, which combines and extends traditional discourse analysis methods. It yields a model of rhetorical metadiscourse, which has allowed us to construct a system of analysis rules that can be implemented in natural language processing software for identifying and highlighting the target expressions. In Section 5, we present the way concept-matching is implemented into a natural language processing software. Finally, in Section 6 we draw some conclusions and set some directions for future work.

2. Rhetorical metadiscourse

The abstract is the showcase of a scientific paper. It should briefly and effectively argue in order to convince the reader of the relevance of the paper with respect to the state of the art. Consequently, the rhetorical principles of abstract writing are highly codified and conventional: an abstract should indicate the author's motivation by presenting the background of the research project described in the paper and point out the problems to be solved. It should also state the author's contribution with respect to the problems, and finally, it should draw conclusions. The author may facilitate the reader's comprehension by announcing these rhetorical functions either by including section titles, like Background, Method, Conclusion, etc., in the abstract –– which is a requirement in some biomedical journals –– or through metadiscourse. We call this kind of metadiscourse rhetorical metadiscourse. If section titles are not available, following Lisacek \textit{et al.} (2005), we propose that automatic natural language processing contribute to the effort of presenting the rhetorical structure of the abstracts by highlighting rhetorical metadiscourse when available\(^3\).

In order to illustrate instances of rhetorical metadiscourse, we present a biomedical research abstract where we have underlined metadiscourse conveying the author's rhetorical strategy. The parts of the text that are not underlined convey scientific knowledge. The expressions underlined are those that our natural language processing system actually highlights in order to facilitate for the reader the interpretation of the scientific propositions, and thus accelerate the evaluation of the article (The expressions underlined by dotted lines are precisions concerning the underlined expressions and are not detected).

\begin{quote}
Most evolutionists agree to consider that our present RNA/DNA/protein world has originated from a simpler world in which RNA played both the role of catalyst and genetic material. Recent findings from structural studies and comparative genomics now allow to get a clearer picture of this transition. These data suggest that evolution occurred in several steps, first from an RNA to an RNA/protein world
\end{quote}

\(^1\) The validity of this assumption is difficult to evaluate. However, the experts actually using the system have confirmed its effectiveness.

\(^2\) The same motivation led a publisher of archeology papers to propose for the first time to present articles following a double articulation, i.e. separating scientific content and rhetorical structure. See Aussenac-Gilles (2006).

\(^3\) This implies that our method does not propose the detection of the rhetorical function of parts of texts when it is not indicated by the author.
(defining two ages of the RNA world) and finally to the present world based on DNA. The DNA world itself probably originated in two steps, first the U-DNA world, following the invention of ribonucleotide reductase, and later on the T-DNA world, with the independent invention of at least two thymidylate synthases. Recently, several authors have suggested that evolution from the RNA world up to the Last Universal Cellular Ancestor (LUCA) could have occurred before the invention of cells. On the contrary, I argue here that evolution of the RNA world taken place in a framework of competing cells and viruses (preys, predators and symbionts). I focus on the RNA-to-DNA transition and expand my previous hypothesis that viruses played a critical role in the emergence of DNA. The hypothesis that DNA and associated mechanisms (replication, repair, recombination) first evolved and diversified in a world of DNA viruses infecting RNA cells readily explains the existence of viral-encoded DNA transaction proteins without cellular homologues. It also potentially explains puzzling observations from comparative genomic, such as the existence of two non-homologous DNA replication machineries in the cellular world. I suggest here a specific scenario for the transfer of DNA from viruses to cells and briefly explore the intriguing possibility that several independent transfers of this kind produced the two cell types (prokaryote/eukaryote) and the three cellular domains presently known (Archaea, Bacteria and Eukarya). (The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. Forterre P. Biochimie. 2005 Sep-Oct; 87(9-10):793-803. Epub 2005 Apr 12.)

Since the author of this abstract employs metadiscourse, the rhetorical structure of this abstract can be articulated by identifying consecutive rhetorical functions that the underlined metadiscourse expressions accomplish. Table 1 summarizes our analysis of this abstract by presenting the rhetorical functions, the metadiscourse conveying them, and the scientific propositions that are positioned by the rhetorical metadiscourse:

<table>
<thead>
<tr>
<th>Rhetorical Function</th>
<th>Metadiscourse</th>
<th>Scientific Proposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>background for research</td>
<td>Most evolutionists...</td>
<td>our present RNA/DNA/protein world has originated from a simpler world...</td>
</tr>
<tr>
<td>background for research</td>
<td>Recent findings...</td>
<td>this transition</td>
</tr>
<tr>
<td>background for research</td>
<td>These data suggest</td>
<td>evolution occurred in several steps...</td>
</tr>
<tr>
<td>background for research</td>
<td>Recently, several authors have suggested</td>
<td>evolution from the RNA world up to the Last Universal Cellular Ancestor (LUCA) could have occurred before the invention of cells</td>
</tr>
<tr>
<td>the author's contribution</td>
<td>On the contrary, I argue here</td>
<td>evolution of the RNA world taken place in a framework of competing cells and viruses</td>
</tr>
<tr>
<td>the author's contribution</td>
<td>I focus on</td>
<td>RNA-to-DNA transition...</td>
</tr>
<tr>
<td>the author's contribution</td>
<td>I... depend my previous hypothesis</td>
<td>viruses played a critical role in the emergence of DNA</td>
</tr>
<tr>
<td>background for research</td>
<td>The hypothesis... readily explains</td>
<td>existence of viral-encoded DNA transaction proteins without cellular homologues</td>
</tr>
<tr>
<td>background for research</td>
<td>This... explains... puzzling observations</td>
<td>the existence of two non-homologous DNA replication machineries in the cellular world</td>
</tr>
<tr>
<td>the author's contribution</td>
<td>I suggest here</td>
<td>a specific scenario for the transfer of DNA from viruses to cells</td>
</tr>
<tr>
<td>the author's contribution</td>
<td>I... explore the... intriguing possibility</td>
<td>several independent transfers of this kind...</td>
</tr>
</tbody>
</table>

Table 1: The analysis of a Medline abstract
In the rest of the paper we will draw on the examples in Table 1 as illustrations of instances of rhetorical metadiscourse. Before describing the model we propose for the representation of rhetorical metadiscourse, we overview in the next section the main methods applied in discourse analysis and show why they are insufficient for modeling these expressions in view of automatic processing.

3. Main methods in discourse analysis

Most studies in discourse analysis start by pointing out that neither is there any consensus in the definition of rhetorical functions, nor is there any common model articulated in linguistic terms for identifying discourse elements that convey rhetorical functions (see e.g. Hyland 2005; Lewin et al. 2001, etc). The reason for this lies in the fact that these discourse elements are determined by content since their role is to accomplish rhetorical and communicative goals. As such they do not correspond to particular lexical or complete syntactic units:

1. As far as the lexicon is concerned, even though scientific discourse is highly standardized, most rhetorical functions are not accomplished by a finite list of lexicalized set expressions.

2. As far as syntax is concerned, the discourse elements like those in Table 1 do not correspond to any particular kinds of complete predicate-argument structures that constitute full propositions, or to any definable components of predicate-argument structures. (cf Recently, several authors have suggested [from the abstract above] vs. previously proposed [from another Medline abstract]. Both expressions convey the rhetorical function of providing background for research, and they are realized in completely different syntactic structures.)

It is specifically their non-propositional character that has been identified by several researchers (Williams 1981; Van de Kopple 1985; Crismore et al. 1983) as the common defining feature of the expressions conveying a particular layer of meaning, which the expressions accomplishing metadiscourse – including what we refer to as rhetorical metadiscourse – also convey.

Due to the fact that particular discourse functions are not conveyed by corresponding lexical and syntactic structures in which they could be mapped in a straightforward way, expressions conveying discourse functions are usually approached from either of two points of view: either through linguistic markers or in conceptual terms. In the following paragraphs we discuss these two approaches, and we point out that neither of them proposes an underlying common model for metadiscourse, which our approach proposes.

3.1. Approach based on lexical markers

The most widespread method in discourse analysis accounts for elements of discourse by characterizing them by lists of markers. This means that whenever a discourse function is accomplished, certain elements are present, and they are considered as markers. Describing discourse functions in terms of linguistic markers has lead to the establishment of various categories of lists of expressions characterized by a mixture of semantic properties, parts of speech, or not characterized by any particular feature. Different authors identify different discourse functions and describe them as being indicated by various markers in the texts (e.g. Van de Kopple 1985; Crismore et al. 1993; Hyland 2005). One of the most influential recent works dealing with metadiscourse is Hyland (2005). In the Appendix (p. 218), the author summarizes the types of metadiscourse he has identified, and defines them as lists of expressions called metadiscourse markers. The rhetorical functions that we have identified above correspond to several metadiscourse markers in Hyland's system, like evidentials (ex. to cite X), announce goals (aim, this section), boosters (always, known), self mention (the
author, our), engagement markers (compare, fairly), etc.

Discourse analysis traditionally serves disciplines related to human sciences, like pedagogy, anthropology, literary studies, etc, and not automatic processing. Thus discourse analysis methods developed in this context need not take into account the fact that there is no one-to-one correspondence between markers and discourse functions, i.e. that the same lexical item does not always accomplish the function of a discourse marker. For automatic processing, however, these marker based analyses are not satisfactory: while the presence of a particular discourse function implies the presence of a set of markers, the opposite is not true, i.e. the presence of a marker does not imply that a particular discourse function is accomplished. For example the word aim is a discourse function marker according to Hyland (2005), but only when used within a longer expression that fulfills the discourse function of “announcing goals”. It cannot be stated that every occurrence of this word is destined to accomplish that discourse function. In sentence (1), aim fulfills this discourse function, but it does not in sentence (2):

1) Strategies aimed at the inhibition of coagulation activation may theoretically be justified.

2) This investigation aimed to characterize the nature of the relationship between the organic and mineral phases ...

As Hyland notes (p218) “Every instance should ... be studied in its sentential co-text.” This, however, has not been done in the framework of traditional marker-based discourse analysis studies, which renders them insufficient for being applied in automatic detection of metadiscourse. Although there are a few discourse markers that invariably signal particular discourse functions, in the vast majority of cases, the presence of a marker does not necessarily signal the accomplishment of a discourse function. Thus without the co-text the discourse function cannot be detected with precision.

Computational approaches to the extraction of metadiscourse also rely on the notion of markers, but in order to yield precise results they need to take into account the co-text. Analysis is carried out by the machine-readable model of metadiscourse constructed by classification algorithms. The model is automatically constructed from the annotation schemes made by human analysts who annotate parts of texts that correspond to pre-defined rhetorical functions. They mark various grammatical and lexical features that they identify as metadiscourse markers. The algorithms associate the sets of markers with rhetorical functions in training corpora, they calculate the correspondences between them, and apply the results for analysis. Such automatically computed algorithms are used for detecting the argumentative structure of scientific discourse or the classification of citation functions (Teufel 1998; Teufel et al. 2002; Teufel et al. 2006; Mizuta et al. 2004). The coverage of automatically constructed discourse models based on annotated examples depends on the coverage of the training examples as well as on the choice of the features. Ideally, such algorithms should provide a model that covers all the instances of discourse, but due to the great variability of the expressions, this is not the case at present. Although such analyzers are effective for detecting recurrent forms, they fail to recognize expressions that do not have a model in the training corpus. Contrary to this data-driven approach, our model described in Section 4 proposes a common underlying structure for metadiscourse based on cognitive analysis.

3.2. Conceptual approach

An alternative approach to discourse functions has been developed in the framework of systemic-functional grammar by Martin (1992). Instead of being conceived of as
compositions of independent markers of various natures, rhetorical functions are generated in system networks where each system is one rhetorical function, and it is composed of semantic concepts. For example the rhetorical function called “claim relevance for research” is fulfilled by a system network composed of two concepts: participants and claims. The example of system network in Table 2 is adapted from Lewin et al. (2001, 33). The words underlined in the example sentence below are parts of the rhetorical function analyzed. “Participants” and “Claim” are the necessary features (concepts) for the realization of the rhetorical function.

Example: Over the past decade, many investigators have called for a closer analysis of the language learning environment of the young infant.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Realization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td></td>
</tr>
<tr>
<td>Phenomena under study</td>
<td>the language learning environment of the young infant</td>
</tr>
<tr>
<td>Research producers</td>
<td></td>
</tr>
<tr>
<td>Products</td>
<td></td>
</tr>
<tr>
<td>Claim</td>
<td></td>
</tr>
<tr>
<td>Magnitude</td>
<td>Over the past decade, many</td>
</tr>
</tbody>
</table>

Table 2: Martin’s conceptual analysis of metadiscourse

This model, contrary to the marker-based approach, does account for co-text, i.e. it includes the entire expression that conveys metadiscourse, and thus it accounts for the conceptual coherence of rhetorical functions. It states that in order to realize the rhetorical function “Claim relevance for research” a rule has to be applied, which requires the presence of a number of participating concepts. The model does not prescribe any other rule. Were we to apply this analysis to automatic processing, as search with a “bag of concepts”, we would detect sentences that contain expressions of participants (phenomena, research producers and research products) and claims (magnitude). However, we would also detect sentences containing the same bag of concepts but not conveying the rhetorical function, like the following sentence:

_The CHAVI investigators_[research producer] _wanted to study_[research product] _those patients who had many_[research producer] _sequential blood samples taken during this plateau in viral load_[phenomenon under study]_.

Our model of rhetorical metadiscourse draws on an alternative approach to analyze discourse: concept matching. As we will describe in the next section, concept-matching links the marker-based and concept-based approaches as well as completes them, and thus proposes a common underlying model of metadiscourse.

4. Concept-matching for modeling metadiscourse

The concept-matching approach for detecting metadiscourse was originally developed in particular text-mining applications (Lisacek et al. 2005; Sandor et al. 2006; Vikef European Project: http://www.vikef.net/; Sandor 2006). Here, we describe concept-matching in general terms by showing how it models the underlying common structure of rhetorical metadiscourse. Let us consider the rhetorical metadiscourse in the abstract in Section 2 that accomplishes the discourse function of yielding background knowledge:
Most evolutionists (agree) to consider
Recent findings … (allow) to get a … picture
These data suggest
Recently, several authors have suggested
(I )… expand … previous hypothesis
The hypothesis … explains
It … explains … observations

In order to model these expressions of rhetorical metadiscourse, we combine, as well as extend the marker-based and the concept-based approaches in the following way:

1. Similarly to Teufel’s (2006) approach, we consider that rhetorical metadiscourse is accomplished by a set of markers. In our model, however, markers are uniquely expressed in terms of concepts, which we call constituent concepts. In this sense our model is similar to the concept-based discourse analysis model of Martin (1992).

2. Contrary to Martin (1992), in our model the same discourse function is not accomplished by exactly the same set of concepts in all of its instantiations. Instead, we have a list of constituent concepts associated with each discourse function, a subset of which is present in the actual expressions. The list of constituent concepts associated with the discourse function of yielding background knowledge is the following:

   a. Background concepts: general, past
   b. Knowledge concepts: researcher, mental activity, object/result of research activity

The various expressions contain at least two of these constituent concepts:

<table>
<thead>
<tr>
<th>background</th>
<th>knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>general</td>
<td>researcher</td>
</tr>
<tr>
<td>past</td>
<td>mental activity</td>
</tr>
<tr>
<td></td>
<td>object/result of research activity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Most evolutionists consider recent findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recently several authors have suggested expand the hypothesis</td>
</tr>
<tr>
<td>[inherited from preceding sentence]</td>
</tr>
<tr>
<td>[inherited from preceding sentence]</td>
</tr>
<tr>
<td>[inherited from preceding sentence]</td>
</tr>
<tr>
<td>[inherited from preceding sentence]</td>
</tr>
</tbody>
</table>

| Table 3: The constituent concepts of background knowledge |

3. Even if they are not full propositions, all of the expressions accomplishing rhetorical...
functions are composed of syntactically linked words, i.e. they are not just a “bag of concepts” as suggested by the concept-based analyses, but syntactically coherent expressions. Syntactic and conceptual coherence are two sides of the same coin: the semantic coherence of rhetorical functions is realized by way of syntactically linking the expressions that convey constituent concepts. The co-text that the various constituent concepts represent with respect to one another is more than simple co-occurrence in the same sentence: it is syntactically linked co-occurrence. This is why, our model includes the syntactic links that actually exist between pairs of constituent concepts in the expressions:

Most evolutionists (agree) to consider

Recent findings... (allow) to get a... picture

These data suggest

Recently, several authors have suggested

(1)... expand (my) previous hypothesis

The hypothesis... explains

It... explains... observations

In summary, concept-matching models rhetorical functions as syntactically linked combinations of realizations of constituent concepts. The notion of constituent concepts yields an underlying general representation for the various lexical units that compose metadiscourse, and thus accounts for their high lexical variety. This is especially important in the case of some constituent concepts whose realizations do not involve a list of synonyms or semantically related words. For example the constituent concept “contrast” (which is a constituent concept of the discourse function of the author’s contribution) is instantiated even within the only example abstract in Section 2 by two different parts-of-speech: adjective (puzzling) and connector (on the contrary), but it can also be an adverb (paradoxically) or a noun (contradiction). On the other hand, the ambiguity of the markers with respect to the rhetorical function they accomplish in the sentences is more effectively cleared up by other syntactically linked markers than by mere co-occurrence.

Figure 1 synthesizes the concept matching model of rhetorical metadiscourse in biomedical abstracts. We identified nine constituent concepts that are altogether involved in rhetorical metadiscourse. The same constituent concept may be involved in several functions (background knowledge, the author’s contribution, etc.). The curved lines represent syntactic dependencies, and the ellipses rhetorical functions. They do not stand, however, for actual syntactic links and subsets. They are used here symbolically to indicate the condition of the presence of syntactic links between two constituent concepts on the one hand, and the fact that the expressions are realized by subsets of syntactically related constituent concepts, on the other hand.

rhetorical metadiscourse in biomedical abstracts
In the next section we describe how this model is used for constructing the natural language processing system that highlights expressions in sentences that convey the rhetorical metadiscourse.

5. Implementation

All of the applications are implemented in the Xerox Incremental Parser (XIP) (Ait et al. 2002), but the concept matching model can be implemented in any natural language processing system that carries out syntactic dependency parsing.

The goal in every application is detecting and highlighting expressions that accomplish the rhetorical functions. The expressions, following the concept-matching model, are syntactically related pairs of words realizing constituent concepts. Syntactic relationships between the words are yielded by the dependency parser, and this is the input for the implementation. It consists in the following three steps:

1. Establishing a list of constituent concepts.

   In order to set up the list of constituent concepts that compose rhetorical metadiscourse we analyze large corpora with various methods. First, we isolate a small number of expressions conveying rhetorical metadiscourse and categorize each word or multi-word expression into a constituent concept. This is done without any predefined algorithm or any other automatic categorization tool. If necessary, during the development of the system, new constituent concepts can be added.

2. Assigning the constituent concepts to a list of words and multi-word expressions.

   The constituent concepts fall into two categories: concepts that are associated with the genre of scientific research literature (researcher, mental activity, object/result of research) and concepts that are more related to general language use (contrast, possibility, general, past, present article).

   Since scientific vocabulary and discourse are relatively uniform, we can compile a fairly representative list of words or expressions that realize each constituent notion relatively rapidly by various ways: using existing lists of scientific vocabulary (like the list of markers in Hyland (2005), conceptual dictionaries (Wordnet) or corpus study (frequency lists).

   Concerning constituent concepts involving general vocabulary, some are mostly realized by restricted language elements, mainly pronouns, and thus are readily available. For example words realizing “the present article” are mainly personal pronouns (I, my, our) or deictic determiners (this). The remaining constituent concepts – “general”, “past”, “potentiality”, “contrast” and “attitude” – neither use restricted language nor are uniquely related to scientific domains. The lists are obtained mainly from studying a corpus of
scientific literature and are incrementally added while working out the system. The coverage of these general language concepts is probably more incomplete than that of the other constituent concepts.

3. Constructing co-occurrence rules.

Co-occurrence rules prescribe which combinations of syntactically linked constituent concepts do accomplish the target rhetorical function. The construction of the rules is carried out experimentally and incrementally with a development corpus. The first hypothesis is that whenever two words realizing constituent concepts are syntactically linked, they convey a rhetorical function. This hypothesis is tested by applying the analysis with this first rule over a corpus. If highlighted expressions do not accomplish the rhetorical function, we impose restrictions. An obvious restriction is that a dependency pair composed of two words realizing the same constituent concept does not accomplish a rhetorical function. Other co-occurrence rules prevent dependency pairs of particular configurations from being highlighted if there is no other dependency pair involved in the sentence. For instance a dependency pair composed of any combination of “general”, “past”, “potentiality”, “contrast” or attitude – the constituent notions that are also parts of general discourse – does not accomplish the rhetorical function. In the following sentence the dependency pair “most[GEN] - possibility[POT]” does not accomplish any rhetorical function since neither of these words is related to another genre-specific word realizing a constituent concept: “Women have the most[GEN] possibility[POT] of suffering from mental disorders during pregnancy and postpartum periods…” The co-occurrence rules aim at establishing the necessary and sufficient conditions for detecting the elements of expressions accomplishing rhetorical functions.

Whereas the performance of highlighting the expressions accomplishing the rhetorical metadiscourse have not yet been evaluated, the other applications that we have mentioned above have, and show satisfactory results in terms of both precision and recall (Lisacek et al. 2005; Sandor 2006).

6. Conclusion

In this paper we have described a natural language processing system that enriches the presentation of biomedical abstracts in the Medline repository by highlighting parts of sentences that accomplish rhetorical metadiscourse. Highlighting focuses the reader's attention, and thus can accelerate his evaluation of the relevance of the abstract for his own purposes.

Our natural language processing system is based on a discourse analysis model, which we call concept-matching. The model provides a representation of discourse elements that accomplish rhetorical functions by joining conceptual and syntactic analyses. Previous analysis models for the descriptions of such discourse elements fail to account for their syntactic coherence. The concept-matching model represents the discourse elements as compositions of syntactically linked expressions realizing constituent concepts that rhetorical metadiscourse is composed of. We have described the application of the concept-matching model biomedical research abstracts. In order to detect these expressions in abstracts in other disciplines, we suppose that the same constituent concepts would be adequate. The vocabulary, however, would need to be adapted.

The concept-matching framework can be implemented using any natural language processing system that carries out dependency parsing. It has previously successfully been applied in various text mining systems using the XIP dependency parser. Its implementation consists in assigning features to lexical items that indicate their conceptual categories and writing co-
occurrence rules that are applied on the output of the parser. The co-occurrence rules regulate which co-occurrences of syntactically linked expressions realizing constituent concepts accomplish the rhetorical function.

We plan some future work in two directions. In general terms, we plan to apply the concept-matching model for detecting other rhetorical functions, and also other elements of discourse like for example expressions that indicate risk or sentiments. Concerning the implementation, we need to work out general methods for easily constructing concept-matching systems for conceptual search.

References


Ágnes Sándor  
Xerox Research Centre Europe  
6. chemin Maupertuis  
38240 Meylan  
France  
Tel.: +33 (0)4 76 61 50 14  
Fax: +33 (0)4 76 61 50 99  
E-mail: Agnes.Sandor@xrce.xerox.com