Discourse and citation analysis with concept-matching

Ágnes Sándor, Aaron Kaplan, Gilbert Rondeau
Xerox Research Centre Europe
6, chemin Maupertuis – 38240 Meylan, France
agnes.sandor@xrce.xerox.com, aaron.kaplan@xrce.xerox.com, gilbert.rondeau@xrce.xerox.com

Abstract:
We present here two natural language processing systems for highlighting passages in scientific texts in order to help researchers to rapidly access relevant knowledge. The first system detects sentences containing expressions fulfilling discourse functions in scientific argumentation like background knowledge, summary sentence, contrast with past findings, etc. The second system detects sentences containing bibliographical references and characterizes the relationship that the authors describe between their work and the work they refer to. The systems are implemented in the Xerox Incremental Parser.

Keywords: discourse functions, citation, concept-matching, robust syntactic parsing.

Résumé:
Nous allons présenter deux outils de traitement automatique de langues naturelles qui surlignent des passages dans des textes scientifiques pour accélérer l’accès aux connaissances. Le premier système détecte des phrases qui contiennent des expressions véhiculant des fonctions discursives dans l’argumentation scientifique comme connaissance de base, phrase-résumé, contraste avec des résultats précédents, etc. Le deuxième système détecte des phrases qui contiennent des références bibliographiques et caractérise la relation décrite par les auteurs entre leur travail et l’œuvre auquel ils se réfèrent. Les systèmes sont implémentés avec le Xerox Incremental Parser.

Mots-clés: fonctions discursives, citation, concept-matching, parsing syntaxique robuste.

1. Introduction

The growing number of scientific research publications makes it difficult for researchers to keep up with the state of the art even in their own domain. Since most research publications are available electronically, natural language processing tools might provide useful support. We propose two tools that are intended to help researchers assimilate the contents of scientific research papers. The first one highlights and types expressions that fulfill relevant discourse functions in scientific argumentation, and the second highlights and types expressions that qualify the relationship between the articles and other articles that they refer to. Both
systems are based on detecting expressions with the concept-matching framework. In sections 2 and 3 we will describe our motivations and the functionalities of both tools. Section 3 explains the concept-matching framework and section 4 our development software and the architecture of the systems.

2. Tool for discourse analysis

Scientific articles are highly structured and follow argumentative patterns that guide the reader in the comprehension of the train of thought described (Hyland 2005, Lewin et al. 2001, Mizuta & Collier 2004, Raveli & Ellis 2004, Teufel 1998, Teufel & Moens 2002, Tognini-Bonelli & Del Lungo Camiciotti 2005). The overall structure of the argumentation is articulated through the formal division of publications into sections, and the finer structure through meta-discourse expressions that make the argumentative discourse functions of the smaller units (sentences or passages) explicit. Often, especially in the domain of experimental research, the titles of the sections are not related to the topics discussed but instead they refer to their discourse functions: introduction, background, methods, result, conclusion, etc. In many domains, these section titles are becoming templates used by a great number of authors, and sometimes even required by the publishers.

However, this formal structuring is insufficient: On the one hand, within one section that is supposed to fulfill the discourse function referred to by its title, the authors very often include digressions fulfilling different discourse functions. For example, a section on results often contains sentences of background knowledge or methods, which also have sections of their own. On the other hand, the diversity of the relevant discourse functions is greater than that of section types. For example an important way of convincing the readers is contrasting one’s results with other results. “Contrast”, however, is not a usual title for a section.

The tool we present marks particular discourse function types of sentences in order to provide the reader with additional support for representing scientific work in a structured way. In its present state our system identifies the following expressions fulfilling relevant discourse functions in scientific argumentation: background knowledge, logical contradiction, an element insufficiently or not known, research trend, summary sentence, contrast with past findings and substantially new finding.

The system has been implemented for processing biomedical literature in the Pubmed repository (Lisacek et al. 2005). The user enters a Pubmed query and an additional list of important keywords that is used for relevance ranking. The output is the list of the retrieved abstracts ranked according to the frequency of the desired keywords, and the sentences containing the above-mentioned content types are highlighted.

2. Tool for citation analysis

Whereas the first tool we presented guides readers in following the train of thought of one article, citation analysis yields help for awareness of “inter-article” relationships. Widely used citation analysis tools are Google Scholar and CiteSeer whose main function is to link citer and citee. Whereas Google Scholar returns a list of publications with the links of the citations, CiteSeer also extracts the passage that includes a reference, and thus indicates its context.

Our tool marks the context of citations according to the type of relationship between citer and cite (Trigg 1983). At its present state the system extracts sentences where the citation is made, and does not consider further sentences that refer to that one, although they might obviously contain important elements. We intend to elaborate wider contexts at a later stage. The system identifies now four kinds of relationships: background knowledge
(general knowledge, knowledge that helps the reader to understand the article or the topic of the article, but that is not linked to the details of the article), based-on (the citing article builds is based in some sense on the article cited, i.e. the cited article has had some effect on the citing article), comparison (the cited article is compared to the citing article (differences or resemblances), but no direct link between the two articles is mentioned, contrary to “based-on”) and assessment (the cited work is assessed, either positively or negatively).

3. Methodology

The discovery of the expressions fulfilling the above-mentioned discourse functions is carried out by the implementation of the concept-matching framework (Sándor 2005). The particular difficulty is the high variability of these expressions both from structural a lexical points of view. In contrast to expressions conveying propositional contents, they do not follow identifiable structural patterns and do not have a single conceptual centre that could serve as an anchor for their identification. The following three sentences illustrate these observations. They all include bibliographic references in order to provide background knowledge:

(1) Semantic Gossiping [3, 4] is a semantic reconciliation method that can be applied to foster semantic interoperability in decentralized settings.

(2) Consequently the necessity of a visual syntax for knowledge representation (KR) languages has been argued frequently in the past [7, 14].

(3) Many other possible approaches to negotiation exist [4], [13].

The relevant expressions conveying the concept “background knowledge” are the following:

(1) Semantic Gossiping [3, 4] is a … method that can be applied.

(2) … has been argued frequently in the past [7, 14].

(3) … other … approaches… exist [4], [13].

In order to establish a common underlying representation of the target expressions, we break down the target concepts into “constituent concepts”. In the case of the above target concept, i.e. “background knowledge”, we have identified three constituent concepts: Previous work[OTHER] provides general[GEN] (background) knowledge[IDEA]. To each constituent concept we assign a list of keywords or expressions. The concept-matching framework is based on the co-occurrence of the expressions of all or a subset of the constituent concepts within the sentences under two types of constraints. The first constraint is the presence of a direct syntactic dependency relationship between pairs of concepts. The second constraint is the application of rules that define the co-occurrence of the subset of the constituent concepts in the sentences that are necessary for matching the target concept. The above sentences are matched due to the fact that the necessary constituent concepts are present and moreover, they are pairwise in syntactic dependency relationships with one another:

(1) DEPENDENCY(Semantic Gossiping[OTHER],is[GEN])
    DEPENDENCY(Semantic Gossiping[OTHER],[3,4][OTHER])
    DEPENDENCY(is[GEN],method[IDEA])
    DEPENDENCY(method[IDEA],can be[GEN])
    DEPENDENCY(can be[GEN],applied[IDEA])

(2) DEPENDENCY(has been[GEN],argued[IDEA])
    DEPENDENCY(argued[IDEA],frequently[GEN])
As for the status of our method among content detection methods, we note that it detects more precise content than search based on bags of words in that it requires the presence of direct syntactic dependencies between classes of keywords. On the other hand, it covers a larger variety of patterns than search based on the detection of precise predicate-argument structures due to two reasons: our keywords in the same class are highly heterogeneous in nature (in the same class we may find verbs, prepositions or adverbs), and matching particular dependency types is not required. We can say that our method is between bag-of-words approaches and bag-of-phrases approaches; we may call it a bag-of-dependency-pairs approach.

4. Development software and architecture

Our systems have been developed with the Xerox Incremental Parser (XIP) (Aït-Mokhtar et al. 2002). XIP is a natural language analysis tool designed for extracting dependency functions between pairs of words within the sentences. The concept-matching grammars are built on top of a general rule-based robust dependency grammar that has been developed in Xerox Research Centre Europe in the XIP formalism. The following schema illustrates the architecture of the system:

![Architecture of the concept-matching systems.](image)

6. Acknowledgement

The development of the tool for citation analysis is funded by the Vikef European project: [http://www.vikef.net/](http://www.vikef.net/).
Bibliography


HYLAND K. (2005), Metadiscourse, Continuum.


