Identifying Strategic Information from Scientific Articles through Sentence Classification

Fidelia Ibekwe-SanJuan[1,2], **Chaomei Chen**[2], **Roberto Pinho**[3]
[1] ELICO - University of Lyon3, France {ibekwe@univ-lyon3.fr}
[2] iSchool, Drexel University, Philadelphia, PA 19104 {cc345@drexel.edu}
[3] University of São Paulo, Brazil {rpinho@icmc.usp.br}

Abstract

We address here the need to assist users in rapidly accessing the most important or strategic information in the text corpus by identifying sentences carrying specific information. More precisely, we want to identify contribution of authors of scientific papers through a categorization of sentences using rhetorical and lexical cues. We built local grammars to annotate sentences in the corpus according to their rhetorical status: objective, new things, results, findings, hypotheses, conclusion, related_word, future work. The annotation is automatically projected automatically onto two other corpora to test their portability across several domains. The local grammars are implemented in the Unitex system. After sentence categorization, the annotated sentences are clustered and users can navigate the result by accessing specific information types. The results can be used for advanced information retrieval purposes.

1. Introduction

We address here the need to assist users in rapidly accessing the most important or strategic information in the text corpus by identifying sentences carrying specific information. More precisely, we want to identify what the author of the paper considers as his/her most important contribution to the research being published. For this, we rely on indicators of scientific discourse structure in order to perform sentence classification, selection and annotation and then visualization from scientific abstracts. Previous studies on scientific discourse (van Dijk 1980 cited in Toefel & Moens 2002) have established that scientific writing can be seen as a problem-solving activity. Thus, it is feasible to present important information from sentences along these dimensions which are almost always present in any scientific writing. More specifically, we seek to identify seven categories of information in abstracts that are interesting for science and technology watch. They are sentences that contain the following: «objective, result, new thing, hypothesis, related work, conclusion, future work». The selected sentences will be annotated with the specific type of information they contain and visualized using techniques from information visualization domain (Lopes *et al.*, 2007; Paulovich *et al.*,2007; Chen 2006).

Our task shares some common points with research on automatic summarization where the goal is to select important sentences to form an abstract. We were particularly inspired by earlier works on the structure of scientific texts (Swales 1990, Salager-Meyer 1992, Teufel & Moens 2002, Orasan 2003). Our overall goal, in line with Teufel & Moens (2002) work, is to highlight the new contribution of the source article and thus situate it with earlier works. This aspect of, "novelty detection" with regard to what is already known has recently become an important focus of the "Document Understanding Conferences" in its 2007 edition (DUC[1]).

To detect automatically sentences in the seven categories listed above, we need to build language resources (LRs). The role of language resources for all human language technologies has long been recognized. However, building language resources can be very labor intensive if these resources are highly domain or corpus-dependent. Hence, it is important to take into consideration maintenance and portability issues. To ensure that our LRs are re-usable on other corpora and easy to maintain, we adopted a surface linguistic analysis using lexico-syntactic patterns that are generic to a given language. We applied these patterns to three different corpora in order to verify their portability across different domains. We also compared and enriched our lexico-syntactic patterns with similar existing systems and with external lexical database such as WordNet. The whole processes leading from sentence annotation to strategic information visualization can be represented by the flowchart in figure 1 hereafter.

2. Corpora

There is a large body of work on the structure of scientific discourse and on cues associated with their communication. Earlier studies have established that the experimental sciences respected more the rhetorical divisions in scientific publications and more often than not, used cues to announce them.

We chose corpora from three different disciplines: experimental science (Quantitative biology), information sciences and astronomy. The first corpus consisted of from preprints on Quantitative biology downloaded from the Open Archives Initiative (OAI[2]) containing the word '*gene*'. 211 publications were found among which we manually studied the first 50 abstracts (written by the authors) in order to design our lexico-syntactic patterns. This first corpus served as a "training corpus" for building the language resources. The patterns were then projected on other test corpora to evaluate their accuracy and

[1] duc.nist.gov/guidelines/2007.html

[2] http://fr.arxiv.org/archive/q-bio

portability. The other two test corpora consisted of: (i) 1000 abstracts of publications from 16 Information Retrieval journals downloaded from the PASCAL[3] database and (ii) a corpus of 1293 abstracts on the Sloan Digital Sky Survey (SDSS[4]) project in astronomy, downloaded from the Thomson ISI[5] database. In the following section, we give details of the patterns derived for each information category.

3. Lexico-syntactic patterns for sentence classification

As observed by previous studies, scientific writing is not a neutral act. It is indeed a social act. Authors have the need to convince their colleagues of the validity of their research, hence they make use of rhetorical cues and a few recurrent patterns (Swales 1990, Teufel & Moens 2002). It is thus feasible to automate the identification of sentences bearing these patterns using information extraction patterns (regular expressions) or templates.

3.1 Levels of Linguistic cues

Meta-discourse patterns found in scientific texts can be distinguished between high-level rhetorical divisions which act as sections announcers and occur at the beginning of a paragraph; and lexico-syntactic patterns which are low-level markers found within the sentences. High-level rhetorical divisions would be the explicit use of a section title such as m*otivation, aim, objective or goal, experiment, results* and *conclusion*. The lower-level patterns are embedded in the rhetorical divisions and introduce more specific information at the sentence level. Both types are frequent in full texts but are also found in abstracts.

3.2 Formalization of patterns

Lexico-syntactic patterns announcing a specific information type are not fixed expressions. They are subject to variations. These variations can occur at different linguistic levels: morphological (gender, number, spelling, inflection), syntactic (active/passive voice, nominal compounding *vs* verbal phrase), lexical (derived form of the same lemma) and semantic (use of synonymous words). The exact surface form of all these variations cannot be known in advance. Hence, categorizing sentences based on these surface patterns requires that we take into account all the places where variations can occur so as to ensure a certain degree of recall. From our manual study of the 50 abstracts in Quantitative biology, we wrote contextual rules in the form of regular expressions implemented as finite state automata in the Unitex[6] system. These automata were then projected on the two test corpora to identify the different categories of sentences. In the table 1 below, we give examples of the some of patterns found for four types of information.

Most patterns necessitate the combined use of POS[7] and lexical information. Figure 2 hereafter shows the automata that identifies "hypothesis" sentences. The grey boxes call other grammar embedded in the current one. The expressive power of such local grammars can be quite high as more simpler grammars can be embedded into more complex ones to achieve a considerable level of complexity.

3.3 Sentence tagging

Once the local grammars have been built and tested, they are used to annotate sentences in the corpus. The annotation is performed automatically using the transducer facility in Unitex. Transducers are variants of the grammars that modify the text by performing a re-writing operation (insert, delete, copy). The information carried by each pattern is inserted at the position where the pattern is found.

3.4. Maintenance and Re-usability of our Lexical resources

We addressed this issue in three ways: 1) applying our patterns across three different corpora to ensure their generality, (2) comparing our patterns with the ones found in the literature, (3) expanding lexical entries in our patterns using external semantic resource such as WordNet. A comparison with Toefel & Moens (2002) on the rhetorical status of sentences in English articles showed that our patterns for certain categories of information (contrast, future work, aim) were in agreement with the ones they found in a corpus of articles in computational linguistics. We also compared our patterns with a system for guided scientific publication writing, called Scipo (Schuster *et al*., 2005) which was also based on the study by Teufel (2002). Scipo is designed to assist non-native English-speaking researchers and students to compose papers in English according to different template structures. One version of this system Scipo Farmácia[8] is tailored for scientific writing in the pharmaceutical domain. Comparison with the patterns in Scipo_Farmacia showed that most of the patterns we had identified were found in Scipo_Farmacia database albeit not always in the same category. Finally, we used WordNet to enrich the list of lexical entries in our patterns (verbs, nouns, adjectives, adverbs) by expanding them with their synonyms. There again, only a small information gain was observed as most of WordNet's synsets were either not appropriate (had different senses than the one meant in scientific writing) or were already in our patterns. Out of the total of 9506 sentences in the SDSS corpus, 1 882 (19%) unique sentences were tagged by our own patterns and 1959 (20%) sentences by the enriched patterns with WordNet. Thus using WordNet did not significantly increase coverage. Hence, we can safely

[3] http://www.inist.fr.
[4] http://www.sdss.org/
[5] Institute for Scientific Information
[6] www-igm.univ-mlv.fr/~unitex/
[7] Part-Of-Speech
[8] http://www.nilc.icmc.usp.br/scipo-farmacia/

assume that our patterns are representative of the meta-discourse structure of most scientific summaries.

4. Visual Analytics for novelty detection and tracking

4.1. Document exploration and sentence annotation

Our overall target is the rapid access to important or strategic information especially as regards new scientific discoveries. The corpus tagged with Unitex is fed into a document clustering system called PEx system (*P*rojection *Ex*plorer) (Paulovich *et a*l., 2007). PEx clusters similar documents and maps them onto a 2D space. The system also builds association rules (Lopes *et al.*, 2007) from regions of similar documents based on user selection. The user can select a group of similar documents and also select a specific information category to visualize on this group (for example "new thing). The systems re-colours the map according to regions rich on the desired information (see figure 3 hereafter). A double-click on a document node take the user to a different window with the underlying annotated abstracts. Each sentence bearing a specific information type underlined and the pattern tags are highlighted using different colors (see figure 4 hereafter).

4.2.Tracking Scientific discoveries

The SDSS (Sloan Digital Sky Survey) project has made a vast amount of observational data available for astronomical research. A primary goal of the NSF-funded project "Coordinated Visualization and Analysis of Sky Survey Data and Astronomical Literature[9]" is to facilitate the understanding of what is known and how it can be related to what is unknown. The sentence classification approach introduced in this paper has the potential of supporting the research in this context. The lexical patterns will enable us to examine all the evidence and discoveries associated with specific hypotheses, and vice versa, group hypotheses together in terms of overlapping evidence. In further research, we will investigate ways of aggregating and synthesizing existing evidence and discoveries in association with a given theory. Another area that the sentence classification may help is the identification of strategic issues concerning the future of a research field. Traditionally, especially in the UK and European countries, scientific foresights are identified based on the input of panels of leading experts. From the literature analysis point of view, a particularly interesting type of sentences or a distinct category of sentences would be ones that specifically address the connections between the main work in a paper and future work. This would provide a complementary source of input regarding the future direction. Sentences identified in this category could be aggregated and synthesized in comparison with foresights developed through panel-based approaches.

[9] http://cluster.cis.drexel.edu/~cchen/projects/sdss/

References

1. Chen C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359-377.
2. Orasan C. (2005) Automatic annotation of Corpora for Text Summarisation: A Comparative Study. In Proceedings of 6th International Conference, CICLing2005, Mexico City, Mexico, February, Springer-Verlag, 670 – 681.
3. Swales J. (1990). Genre Analysis: English in academic and research settings, Cambridge University Press, 1990.
4. Salanger-Meyer F. (1990) Discoursal movements in medical English abstracts and their linguistic exponents: a genre analysis study, INTERFACE: *Journal of Applied Linguistics* 4(2), 1990, 107 - 124
5. Teufel S., Moens M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status, Computational Linguistics, 2002,vol. 28(4), 409-445.
6. Lopes A. A., Pinho R., Paulovich F. V., Minghim R.. Visual text mining using association rules, *Computers & Graphics*, v. 31, p. 316-326, 2007.
7. Paulovich F. V., Oliveira M. C. F., Minghim R., The projection explorer: A flexible tool for projection-based multidimensional visualization. *In*: XX Brazilian Symposium on Computer Graphics and Image Processing SIBIGRAPI 2007, Belo Horizonte, IEEE Computer Society Press, 2007.

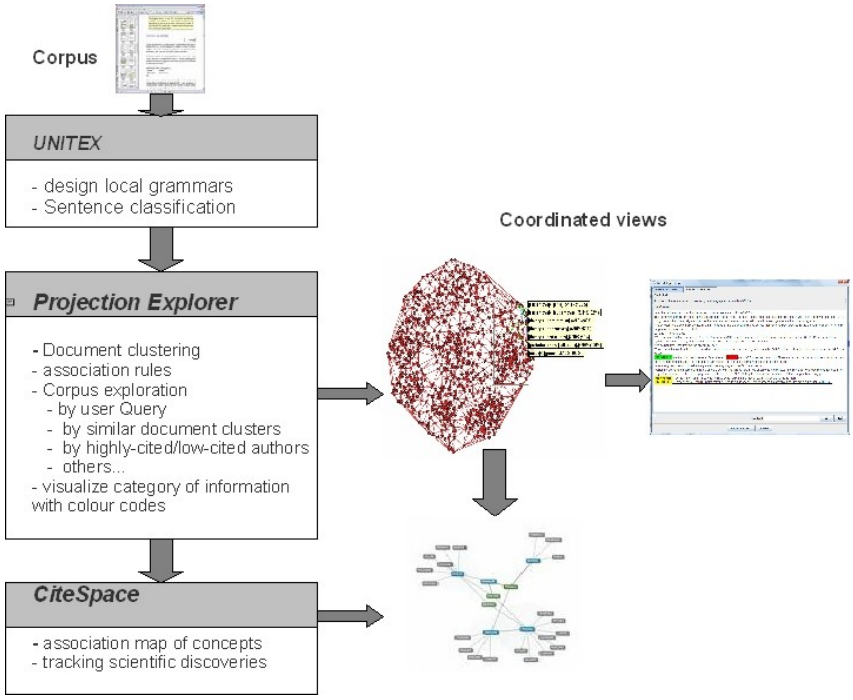


Figure 1. Flowchart of the different tools and processes involved in strategic information visualization.

Information type	Patterns
OBJECTIVE	In this_{article|paper|study|research|work}... We_{examine|investigate|describe|present|outline|introduce|consider}.... DET_{motivation:|aim|goal|objective}...
NEWTHING	Here, we propose a novel (...) approach... This analysis reveals... Emerging evidence suggests that... Interestingly, our results indicate that...
HYPOTHESIS	DET_NP_{may|might}_{ADV|V_NP}... Our findings support the view that... DET_NP_can_{V|NP.}..
FUTURE_WORK	{Further|Future|more}_{work|investigation|observation}_<verb>...

Table 1. Some example of lexico-syntactic patterns announcing specific information category

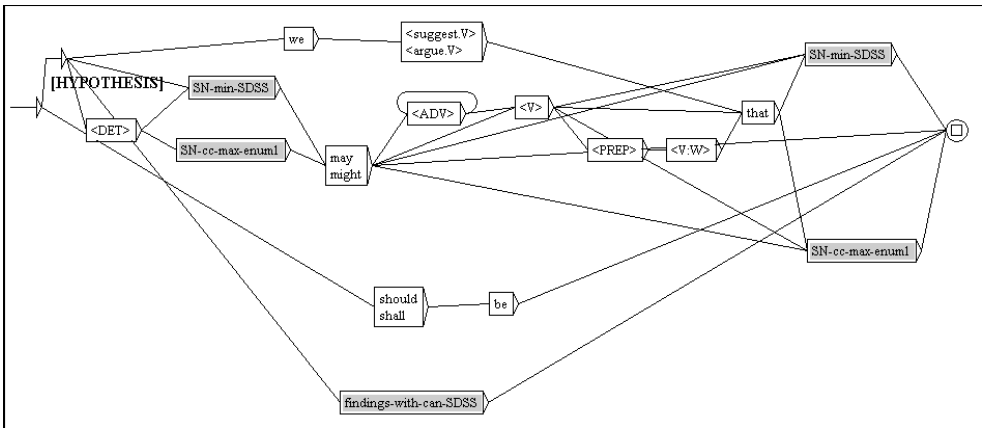


Figure 2. Finite state automata identifying sentences announcing the 'hypothesis"

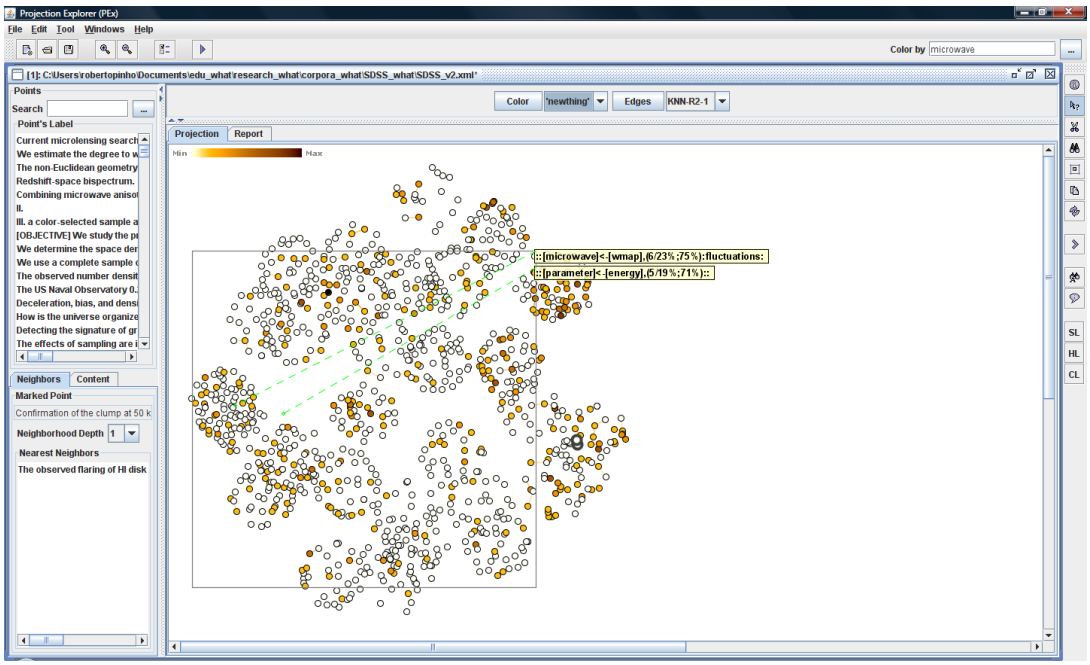
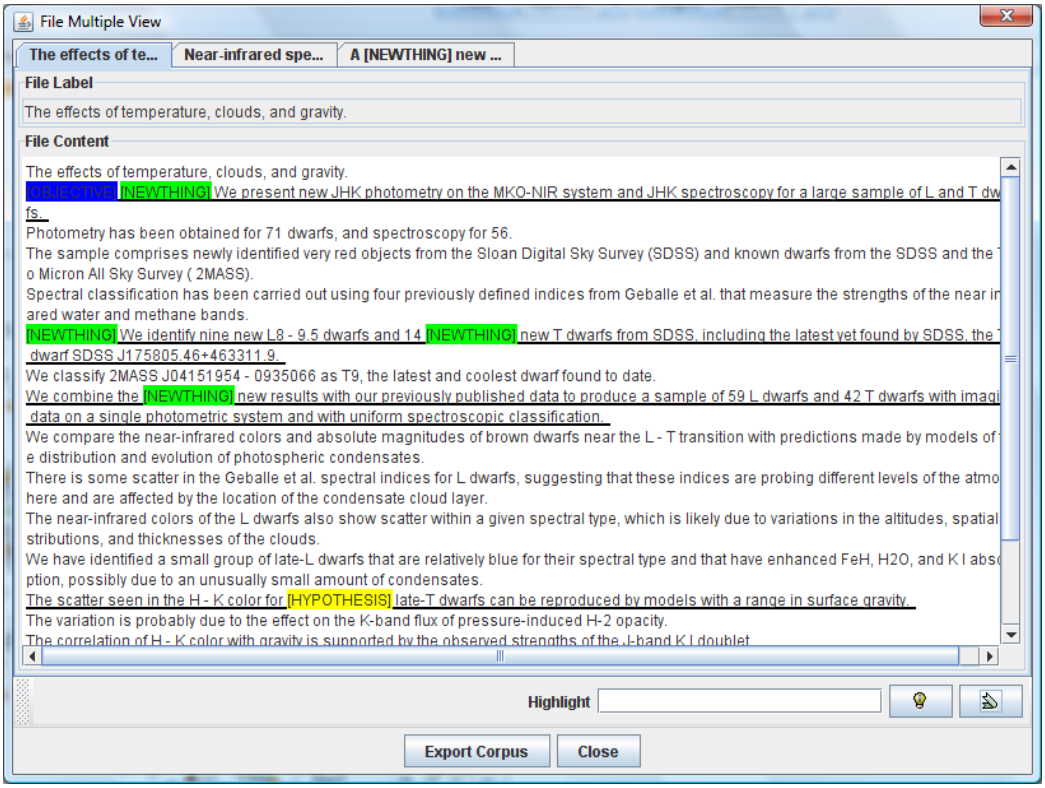


Figure 3. Map of similar documents in the SDSS corpus.
Selected documents (black nodes) are those with "*new thing*" tag in their abstracts.



4. Annotated and highlighted abstract with color codes with the PEx system.