

A Design Methodology for a Biomedical Literature Indexing Tool Using the Rhetoric of Science

Robert E. Mercer

University of Western Ontario,
London, Ontario, N6A 5B7
mercer@csd.uwo.ca

Chrysanne Di Marco

University of Waterloo,
Waterloo, Ontario, N2L 3G1
cdimarco@uwaterloo.ca

Abstract

Literature indexing tools provide researchers with a means to navigate through the network of scholarly scientific articles in a subject domain. We propose that more effective indexing tools may be designed using the links between articles provided by citations.

With the explosion in the amount of scientific literature and with the advent of artifacts requiring more sophisticated indexing, a means to provide more information about the citation relation in order to give more intelligent control to the navigation process is warranted. In order to navigate a citation index in this more-sophisticated manner, the citation index must provide not only the citation-link information, but also must indicate the function of the citation. The design methodology of an indexing tool for scholarly biomedical literature which uses the rhetorical context surrounding the citation to provide the citation function is presented. In particular, we discuss how the scientific method is reflected in scientific writing and how this knowledge can be used to decide the purpose of a citation.

within a research field by linking together works whose methods and results are in some way mutually relevant. Customarily, authors include citations in their papers to indicate works that are foundational in their field, background for their own work, or representative of complementary or contradictory research. Another researcher may then use the presence of citations to locate articles she needs to know about when entering a new field or to read in order to keep track of progress in a field where she is already well-established. But, with the explosion in the amount of scientific literature, a means to provide more information in order to give more intelligent control to the navigation process is warranted. A user normally wants to navigate more purposefully than "Find all articles citing a source article". Rather, the user may wish to know whether other experiments have used similar techniques to those used in the source article, or whether other works have reported conflicting experimental results. In order to navigate a citation index in this more-sophisticated manner, the citation index must contain not only the citation-link information, but also must indicate the function of the citation in the citing article.

The goal of our research project is the design and implementation of an indexing tool for scholarly biomedical literature which uses the text surrounding the citation to provide information about the binary relation between the two papers connected by a citation. In particular, we are interested in how the scientific method structures the way in which ideas, results, theories, etc. are presented in scientific writing and how the style of presentation indicates the purpose of citations, that is, what the relationship is between the cited and citing papers.

Our interest in the connection between scientific literature (our focus), ontologies, and databases is that the content and structure of each of these three repositories of scientific knowledge has its foundations in the method of science. Our purpose here is twofold: to make explicit our design methodology for an indexing tool that uses

1 Introduction

1.1 The aim of citation indexing

Indexing tools, such as CiteSeer (Bollacker et al., 1999), play an important role in the scientific endeavour by providing researchers with a means to navigate through the network of scholarly scientific papers using the connections provided by citations. Citations relate articles

the rhetoric of science as its foundation to see whether the ideas that underly our methodology can cross-fertilize the enquiry into the other two areas, and to discuss the tool itself with the purpose of making known that there exists a working tool which can assist the development of other projects.

A *citation* may be formally defined as a portion of a sentence in a citing document which references another document or a set of other documents collectively. For example, in sentence 1 below, there are two citations: the first citation is *Although the 3-D structure... progress*, with the set of references (Eger et al., 1994; Kelly, 1994); the second citation is *it was shown... submasses* with the single reference (Coughlan et al., 1986).

- (1) Although the 3-D structure analysis by x-ray crystallography is still in progress (Eger et al., 1994; Kelly, 1994), it was shown by electron microscopy that XO consists of three submasses (Coughlan et al., 1986).

A *citation index* enables efficient retrieval of documents from a large collection—a citation index consists of source items and their corresponding lists of bibliographic descriptions of citing works. The use of citation indexing of scientific articles was invented by Dr. Eugene Garfield in the 1950s as a result of studies on problems of medical information retrieval and indexing of biomedical literature. Dr. Garfield later founded the Institute for Scientific Information (ISI), whose Science Citation Index (Garfield, no date) is now one of the most popular citation indexes. Recently, with the advent of digital libraries, Web-based indexing systems have begun to appear (e.g., ISI's 'Web of Knowledge', CiteSeer (Bollacker et al., 1999)).

Authors of scientific papers normally include citations in their papers to indicate works that are connected in an important way to their paper. Thus, a citation connecting the source document and a citing document serves one of many functions. For example, one function is that the citing work gives some form of credit to the work reported in the source article. Another function is to criticize previous work. Other functions include: foundational works in their field, background for their own work, works which are representative of complementary or contradictory research.

The aim of citation analysis studies has been to categorize and, ultimately, to classify the function of scientific citations automatically. Many citation classification schemes have been developed, with great variance in the number and nature of categories used. Garfield (1965) was the first to define a classification scheme, while Finney (1979) was the first to suggest that a citation classifier could be automated. Other classification schemes include those by Cole (1975), Duncan, Anderson, and

McAleese (1981), Frost (1979), Lipetz (1965), Moravcsik and Murugesan (1975), Peritz (1983), Small (1978), Spiegel-Rösing (1977), and Weinstock (1971). Within this representative group of classification schemes, the number of categories ranges from four to 26. Examples of these categories include a *contrastive*, *supportive*, or *corrective* relationship between citing and cited works. But, the author's purpose for including a citation is not apparent in the citation per se. Determining the nature of the exact relationship between a citing and cited paper often requires some level of understanding of the text in which the citation is embedded.

1.2 Citation indexing in biomedical literature analysis

In the biomedical field, we believe that the usefulness of automated citation classification in literature indexing can be found in both the larger context of managing entire databases of scientific articles or for specific information-extraction problems. On the larger scale, database curators need accurate and efficient methods for building new collections by retrieving articles on the same topic from huge general databases. Simple systems (e.g., (Andrade and Valencia, 1998), (Marcotte et al., 2001)) consider only keyword frequencies in measuring article similarity. More-sophisticated systems, such as the Neighbors utility (Wilbur and Coffee, 1994), may be able to locate articles that appear to be related in *some* way (e.g., finding related Medline abstracts for a set of protein names (Blaschke et al., 1999)), but the lack of specific information about the nature and validity of the relationship between articles may still make the resulting collection a less-than-ideal resource for subsequent analysis. Citation classification to indicate the nature of the relationships between articles in a database would make the task of building collections of related articles both easier and more accurate. And, the existence of additional knowledge about the nature of the linkages between articles would greatly enhance navigation among a space of documents to retrieve meaningful information about the related content.

A specific problem in information extraction that may benefit from the use of citation categorization involves mining the literature for protein-protein interactions (e.g., (Blaschke et al., 1999), (Marcotte et al., 2001), (Thomas et al., 2000)). Currently, even the most-sophisticated systems are not yet capable of dealing with all the difficult problems of resolving ambiguities and detecting hidden knowledge. For example, Blaschke et al.'s system (1999) is able to handle fairly complex problems in detecting protein-protein interactions, including constructing the network of protein interactions in cell-cycle control, but important implicit knowledge is not recognized. In the case of cell-cycle analysis for *Drosophila*, their system is able to determine that relationships exist between **Cak**,

Cdk7, **CycH**, and **Cdk2**: **Cak** inhibits/phosphorylates **Cdk7**, **Cak** activates/phosphorylates **Cdk2**, **Cdk7** phosphorylates **Cdk2**, **CycH** phosphorylates **Cak** and **CycH** phosphorylates **Cdk2**. However, the system is not able to detect that **Cak** is actually a complex formed by **Cdk7** and **CycH**, and that the **Cak** complex regulates **Cdk2**. While the earlier literature describes inter-relationships among these proteins, the recognition of the generalization in their structure, i.e., that these proteins are part of a complex, is contained only in more-recent articles: “There is an element of generalization implicit in later publications, embodying previous, more dispersed findings. A clear improvement here would be the generation of associated weights for texts according to their level of generality” (Blaschke et al., 1999). Citation categorization could provide just these kind of ‘ancestral’ relationships between articles—whether an article is foundational in the field or builds directly on closely related work—and, if automated, could be used in forming collections of articles for study that are labelled with explicit semantic and rhetorical links to one another. Such collections of semantically linked articles might then be used as ‘thematic’ document clusters (cf. Wilbur (2002)) to elicit much more meaningful information from documents known to be closely related.

An added benefit of having citation categories available in text corpora used for studies such as extracting protein-protein interactions is that more, and more-meaningful, information may be obtained. In a potential application for our research, Blaschke et al. (1999) noted that they were able to discover many more protein-protein interactions when including in the corpus those articles found to be related by the Neighbors facility (Wilbur and Coffee, 1994) (285 versus only 28 when relevant protein names alone were used in building the corpus). Lastly, very difficult problems in scientific and biomedical information extraction that involve aspects of deep-linguistic meaning may be resolved through the availability of citation categorization in curated texts: synonym detection, for example, may be enhanced if different names for the same entity occur in articles that can be recognized as being closely related in the scientific research process.

2 Our Guiding Principles

2.1 Scientific writing and the rhetoric of science

The automated labelling of citations with a specific citation function requires an analysis of the linguistic features in the text surrounding the citation, coupled with a knowledge of the author’s pragmatic intent in placing the citation at that point in the text. The author’s purpose for including citations in a research article reflects the fact that researchers wish to communicate their results to their scientific community in such a way that their re-

sults, or *knowledge claims*, become accepted as part of the body of scientific knowledge. This persuasive nature of the scientific research article, how it contributes to making and justifying a knowledge claim, is recognized as the defining property of scientific writing by rhetoricians of science, e.g., (Gross, 1996), (Gross et al., 2002), (Hyland, 1998), (Myers, 1991). Style (lexical and syntactic choice), presentation (organization of the text and display of the data), and argumentation structure are noted as the rhetorical means by which authors build a convincing case for their results.

Our approach to automated citation classification is based on the detection of fine-grained linguistics cues in scientific articles that help to communicate these rhetorical stances and thereby map to the pragmatic purpose of citations. As part of our overall research methodology, our goal is to map the various types of pragmatic cues in scientific articles to rhetorical meaning. Our previous work has described the importance of *discourse cues* in enhancing inter-article cohesion signalled by citation usage (Mercer and Di Marco, 2003), (Di Marco and Mercer, 2003). We have also been investigating another class of pragmatic cues, *hedging cues*, (Mercer, Di Marco, and Kroon, 2004), that are deeply involved in creating the pragmatic effects that contribute to the author’s knowledge claim by linking together a mutually supportive network of researchers within a scientific community.

2.2 Results of our previous studies

In our preliminary study (Mercer and Di Marco, 2003), we analyzed the frequency of the cue phrases from (Marcu, 1997) in a set of scholarly scientific articles. We reported strong evidence that these cue phrases are used in the citation sentences and the surrounding text with the same frequency as in the article as a whole. In subsequent work (Di Marco and Mercer, 2003), we analyzed the same dataset of articles to begin to catalogue the fine-grained discourse cues that exist in citation contexts. This study confirmed that authors do indeed have a rich set of linguistic and non-linguistic methods to establish discourse cues in citation contexts.

Another type of linguistic cue that we are studying is related to hedging effects in scientific writing that are used by an author to modify the affect of a ‘knowledge claim’. Hedging in scientific writing has been extensively studied by Hyland (1998), including cataloging the pragmatic functions of the various types of hedging cues. As Hyland (1998) explains, “[Hedging] has subsequently been applied to the linguistic devices used to qualify a speaker’s confidence in the truth of a proposition, the kind of caveats like *I think*, *perhaps*, *might*, and *maybe* which we routinely add to our statements to avoid commitment to categorical assertions. Hedges therefore express tentativeness and possibility in communication, and their ap-

appropriate use in scientific discourse is critical (p. 1)”.

The following examples illustrate some of the ways in which hedging may be used to deliberately convey an attitude of uncertainty or qualification. In the first example, the use of the verb *suggested* hints at the author’s hesitancy to declare the absolute certainty of the claim:

- (2) The functional significance of this modulation is suggested by the reported inhibition of MeSo-induced differentiation in mouse erythroleukemia cells constitutively expressing c-myb.

In the second example, the syntactic structure of the sentence, a fronted adverbial clause, emphasizes the effect of qualification through the rhetorical cue *Although*. The subsequent phrase, *a certain degree*, is a lexical modifier that also serves to limit the scope of the result:

- (3) Although many neuroblastoma cell lines show a certain degree of heterogeneity in terms of neurotransmitter expression and differentiative potential, each cell has a prevalent behavior in response to differentiation inducers.

In Mercer (2004), we showed that the hedging cues proposed by Hyland occur more frequently in citation contexts than in the text as a whole. With this information we conjecture that hedging cues are an important aspect of the rhetorical relations found in citation contexts and that the pragmatics of hedges may help in determining the purpose of citations.

We investigated this hypothesis by doing a frequency analysis of hedging cues in citation contexts in a corpus of 985 biology articles. We obtained statistically significant results (summarized in Table 1 indicating that hedging is used more frequently in citation contexts than the text as a whole. Given the presumption that writers make stylistic and rhetorical choices purposefully, we propose that we have further evidence that connections between fine-grained linguistic cues and rhetorical relations exist in citation contexts.

Table 1 shows the proportions of the various types of sentences that contain hedging cues, broken down by hedging-cue category (verb or nonverb cues), according to the different sections in the articles (background, methods, results and discussion, conclusions). For all but one combination, citation sentences are more likely to contain hedging cues than would be expected from the overall frequency of hedge sentences ($p \leq .01$). Citation ‘window’ sentences (i.e., sentences in the text close to a citation) generally are also significantly ($p \leq .01$) more likely to contain hedging cues than expected, though for certain combinations (methods, verbs and nonverbs; res+disc, verbs) the difference was not significant.

Tables 2, 3, and 4 summarize the occurrence of hedging cues in citation ‘contexts’ (a citation sentence and the

surrounding citation window). Table 5 shows the proportion of hedge sentences that either contain a citation, or fall within a citation window; Table 5 suggests (last 3-column column) that the proportion of hedge sentences containing citations or being part of citation windows is at least as great as what would be expected just by the distribution of citation sentences and citation windows.

Table 1 indicates (statistically significant) that in most cases the proportion of hedge sentences in the citation contexts is greater than what would be expected by the distribution of hedge sentences. Taken together, these conditional probabilities support the conjecture that hedging cues and citation contexts correlate strongly. Hyland (1998) has catalogued a variety of pragmatic uses of hedging cues, so it is reasonable to speculate that these uses can be mapped to the rhetorical meaning of the text surrounding a citation, and from thence to the function of the citation.

3 Our Design Methodology

3.1 The Tool

The indexing tool that we are designing enhances a standard citation index by labelling each citation with the function of that citation. That is, given an agreed-upon set of citation functions, our tool will categorize a citation automatically into one of these functional categories. To accomplish this automatic categorization we are using a decision tree: given a set of features, which combinations of features map to which citation function. Our current focus is the biomedical literature, but we are certain that our tool can be used for the experimental sciences. We are not certain whether the tool can be generalized beyond this corpus (Frost, 1979).

In the following we describe in more detail the three aspects of our design methodology: the research program, the tool implementation, and its evaluation. Our basic assumption is that citations form links to other articles for much the same purpose and in much the same way as links to other parts of the same article. These intra-textual and inter-textual linkages are made to create a coherent presentation to convince the reader that the content of the article is of value. The presentation is made cohesive by use of linguistic and stylistic devices that have been catalogued by rhetoricians and which we believe may be detected by automated means.

The research program will

- develop a catalogue of linguistic and non-linguistic cues that capture both the linguistic and stylistic techniques as well as the extensive body of knowledge that has accumulated about the rhetoric of science and how science is written about;

Table 1: Proportion of sentences containing hedging cues, by type of sentence and hedging cue category.

	Verb Cues			Nonverb Cues			All Cues		
	Cite	Wind	All	Cite	Wind	All	Cite	Wind	All
background	0.15	0.11	0.13	0.13	0.13	0.12	0.25	0.22	0.24
methods	0.09	0.06	0.06	0.05	0.04	0.04	0.14	0.10	0.09
res+disc	0.22	0.16	0.16	0.15	0.14	0.14	0.32	0.27	0.27
conclusions	0.29	0.22	0.20	0.18	0.19	0.15	0.42	0.36	0.32

Table 2: Number and proportion of citation contexts containing a hedging cue, by section and location of hedging cue.

	Contexts		Sentences		Windows	
	#	%	#	%	#	%
background	3361	0.33	2575	0.25	2679	0.26
methods	1089	0.18	801	0.14	545	0.09
res+disc	7257	0.44	5366	0.32	4660	0.28
conclusions	338	0.58	245	0.42	221	0.38

- develop computationally realizable methods to detect these cues;
- connect these cues to rhetorical relations; and
- organize the knowledge that these rhetorical relations represent as features in a decision tree that produces the intended function of the citation.

Our purpose in using a decision tree is three-fold. Firstly, the decision tree gives us ready access to the citation-function decision rules. Secondly, we aim to have a working indexing tool whenever we add more knowledge to the categorization process. This goal appears very feasible given our design choice to use a decision tree: adding more knowledge only refines the decision-making procedure of the previous version. And thirdly, as we gain more experience (currently, we are building the decision tree by hand), we intend to use machine learning techniques to enhance our tool by inducing a decision tree.

3.2 The Research Program

Our basic assumption is that the rhetorical relations that will provide the information that will allow the tool to categorize the citations in a biomedical article are evident to the reader through the use of surface linguistic cues, cues which are linguistically-based but require some knowledge that is not directly derivable from the text, and some cues which are known to the culture of scientific readers-writers because of the practice of science and how this practice influences communication through the writing.

We rely on the notion that rhetorical information is realized in linguistic ‘cues’ in the text, some of which,

although not all, are evident in surface features (cf. Hyland (1998) on surface hedging cues in scientific writing). Since we anticipate that many such cues will map to the same rhetorical features that give evidence of the text’s argumentative and pragmatic meaning, and that the interaction of these cues will likely influence the text’s overall rhetorical effect, the formal *rhetorical relation* (cf. (Mann and Thompson, 1988)) appears to be the appropriate feature for the basis of the decision tree. So, our long-term goal is to map between the textual cues and rhetorical relations. Having noted that many of the cue words in the prototype are discourse cues, and with two recent important works linking discourse cues and rhetorical relations ((Knott, 1996; Marcu, 1997)), we began our investigation of this mapping with discourse cues. We have some early results that show that discourse cues are used extensively with citations and that some cues appear much more frequently in the citation context than in the full text (Mercer and Di Marco, 2003). Another textual device is the hedging cue, which we are currently investigating (Mercer, Di Marco, and Kroon, 2004).

Although our current efforts focus on cue words which are connected to organizational effects (discourse cues), and writer intent (hedging cues), we are also interested in other types of cues that are associated more closely to the purpose and method of science. For example, the scientific method is, more or less, to establish a link to previous work, set up an experiment to test an hypothesis, perform the experiment, make observations, then finally compile and discuss the importance of the results of the experiment. Scientific writing reflects this scientific method and its purpose: one may find evidence even at the coarsest granularity of the IMRaD structure in scientific articles. At a finer granularity, we have many target-

Table 3: Proportion of citation contexts containing a verbal hedging cue, by section and location of hedging cue.

	Contexts		Sentences		Windows	
	#	%	#	%	#	%
background	1967	0.19	1511	0.15	1479	0.15
methods	726	0.12	541	0.09	369	0.06
res+disc	4858	0.29	3572	0.22	2881	0.17
conclusions	227	0.39	168	0.29	139	0.24

Table 4: Proportion of citation contexts containing a nonverb hedging cue, by section and location of hedging cue.

	Contexts		Sentences		Windows	
	#	%	#	%	#	%
background	1862	0.18	1302	0.13	1486	0.15
methods	432	0.07	295	0.05	198	0.03
res+disc	3751	0.23	2484	0.15	2353	0.14
conclusions	186	0.32	107	0.18	111	0.19

ted words to convey the notions of procedure, observation, reporting, supporting, explaining, refining, contradicting, etc. More specifically, science categorizes into taxonomies or creates polarities. Scientific writing then tends to compare and contrast or refine. Not surprisingly, the morphology of scientific terminology exhibits comparison and contrasting features, for example, *exo-* and *endo-*. Science needs to measure, so scientific writing contains measurement cues by referring to scales (*0–100*), or using comparatives (*larger*, *brighter*, etc.). Experiments are described as a sequence of steps, so this is an implicit method cue.

Since the inception of the formal scientific article in the seventeenth century, the process of scientific discovery has been inextricably linked with the actions of writing and publishing the results of research. Rhetoricians of science have gradually moved from a purely descriptive characterization of the science genre to full-fledged field studies detailing the evolution of the scientific article. During the first generation of rhetoricians of science, e.g., (Myers, 1991), (Gross, 1996), (Fahnestock, 1999), the persuasive nature of the scientific article, how it contributes to making and justifying a knowledge claim, was recognized as the defining property of scientific writing. Style (lexical and syntactic choice), presentation (organization of the text and display of the data), and argumentation structure were noted as the rhetorical means by which authors build a convincing case for their results. Recently, second-generation rhetoricians of science (e.g., (Hyland, 1998), (Gross et al., 2002)) have begun to methodically analyze large corpora of scientific texts with the purpose of cataloguing specific stylistic and rhetorical features that are used to create the pragmatic effects that contribute to the author’s knowledge claim. One particu-

lar type of pragmatic effect, *hedging*, is especially common in scientific writing and can be realized through a wide variety of linguistic choices.

To catalogue these cues and to propose a mapping from these cues to rhetorical relations, we suggest a research program that consists of two phases. One phase is theory-based: we are applying our knowledge from computational linguistics and the rhetoric of science to develop a set of principles that guide the development of rules. Another phase is data-driven. This phase will use machine-learning techniques to induce a decision tree.

Our two approaches are guided by a number of factors. Firstly, the initial set of 35 categories ((Garzone, 1996), (Garzone and Mercer, 2000)) were developed by combining and adding to the previous work from the information science community with a preliminary manual study of citations in biochemistry and physics articles. Secondly, our next stages, cataloguing linguistic cues, will require manual work by rhetoricians. Thirdly, and perhaps most importantly, one group of cues is not found in the text, but is rather a set of cultural guidelines that are accepted by the scientific community for which the article is being written. Lastly, we are interested not in the connection between the citation functions and these cues per se, but rather the citation functions and the rhetorical relations that are signalled by the cues.

3.3 The Tool Implementation

Concerning the features on which the decision tree makes its decisions, we have started with a simple, yet fully automatic prototype (Garzone, 1996) which takes journal articles as input and classifies every citation found therein. Its decision tree is very shallow, using only sets of cue-words and polarity switching words (*not*, *however*,

Table 5: Proportion of hedge sentences that contain citations or are part of a citation window, by section and hedging cue category.

	Verb Cues			Nonverb Cues			All Cues		
	Cite	Wind	None	Cite	Wind	None	Cite	Wind	None
background	0.52	0.23	0.25	0.47	0.28	0.25	0.49	0.26	0.26
methods	0.25	0.16	0.59	0.20	0.15	0.65	0.23	0.16	0.61
res+disc	0.26	0.19	0.55	0.21	0.19	0.60	0.23	0.19	0.58
conclusions	0.16	0.14	0.70	0.14	0.16	0.70	0.15	0.14	0.71

etc.), some simple knowledge about the IMRaD structure¹ of the article together with some simple syntactic structure of the citation-containing sentence. The prototype uses 35 citation categories. In addition to having a design which allows for easy incorporation of more-sophisticated knowledge, it also gives flexibility to the tool: categories can be easily coalesced to give users a tool that can be tailored to a variety of uses.

Although we anticipate some small changes to the number of categories due to category refinement, the major modifications to the decision tree will be driven by a more-sophisticated set of features associated with each citation. When investigating a finer granularity of the IMRaD structure, we came to realize that the structure of scientific writing at all levels of granularity was founded on *rhetoric*, which involves both argumentation structure as well as stylistic choices of words and syntax. This was the motivation for choosing the rhetoric of science as our guiding principle.

3.4 Evaluation of the Tool

Finally, as for our prototype system, at each stage of development the tool will be evaluated:

- A test set of citations will be developed and will be initially manually categorized by humans knowledgeable in the scientific field that the articles represent.
- Of most essential interest, the classification accuracy of the citation-indexing tool will be evaluated: we propose to use a combination of statistical testing and validation by human experts.
- In addition, we would like to assess the tool’s utility in real-world applications such as database curation for studies in biomedical literature analysis. We have suggested earlier that there may be many uses of this tool, so a significant aspect of the value of our tool will be its ability to enhance other research projects.

¹The corpus of biomedical papers all have the standard Introduction, Methods, Results, and Discussion or a slightly modified version in which Results and Discussion are merged.

4 Conclusions and Future Work

The purposeful nature of citation function is a feature of scientific writing which can be exploited in a variety of ways. We anticipate more-informative citation indexes as well as more-intelligent database curation. Additionally, sophisticated information extraction may be enhanced when better selection of the dataset is enabled. For example, synonym detection in a corpus of papers may be made more tractable when the corpus is comprised of related papers derived from navigating a space of linked citations.

In this paper we have motivated our approach to developing a literature-indexing tool that computes the functions of citations. We have proposed that the function of a citation may be determined by analyzing the rhetorical intent of the text that surrounds it. This analysis is founded on the guiding principle that the scientific method is intrinsic to scientific writing.

Our early investigations have determined that linguistic cues and citations are related in important ways. Our future work will be to map these linguistic cues to rhetorical relations and other pragmatic functions so that this information can then be used to determine the purpose of citations.

Acknowledgements

We thank Mark Garzone and Fred Kroon for their participation in this project. Our research has been financially supported by the Natural Sciences and Engineering Research Council of Canada and by the Universities of Western Ontario and Waterloo.

References

- Miguel A. Andrade and Alfonso Valencia. 1998. Automatic Extraction of Keywords from Scientific Text: Application to the Knowledge Domain of Protein Families. *Bioinformatics*, 14(7):600–607.
- Christian Blaschke, Miguel A. Andrade, Christos Ouzounis, and Alfonso Valencia. 1999. Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions. *International Conference*

- on *Intelligent Systems for Molecular Biology (ISMB 1999)*, 60–67.
- B. Bollacker, S. Lawrence, and C.L. Giles. 1999. A System for Automatic Personalized Tracking of Scientific Literature on the Web. In *Digital Libraries 99—The Fourth ACM Conference on Digital Libraries*, 105–113. ACM Press, New York.
- S. Cole. 1975. The Growth of Scientific Knowledge: Theories of Deviance as a Case Study. In *The Idea of Social Structure: Papers in Honor of Robert K. Merton*, 175–220. Harcourt Brace Jovanovich, New York.
- Chrysanne Di Marco and Robert E. Mercer. 2003. Toward a Catalogue of Citation-related Rhetorical Cues in Scientific Texts. In *Proceedings of the Pacific Association for Computational Linguistics (PACLING 2003) Conference*. Halifax, Canada, August 2003.
- E.B. Duncan, F.D. Anderson, and R. McAleese. 1981. Qualified Citation Indexing: Its Relevance to Educational Technology. In *Information Retrieval in Educational Technology: Proceedings of the First Symposium on Information Retrieval in Educational Technology*, 70–79. University of Aberdeen.
- Jeanne Fahnestock. 1999. *Rhetorical Figures in Science*. Oxford University Press.
- B. Finney. 1979. The Reference Characteristics of Scientific Texts. Master's thesis, The City University of London.
- C. Frost. 1979. The Use of Citations in Literary Research: A Preliminary Classification of Citation Functions. *Library Quarterly*, 49:399–414.
- Eugene Garfield. 1965. Can Citation Indexing Be automated? In M.E. Stevens et al., editors, *Statistical Association Methods for Mechanical Documentation (NBS Misc. Pub. 269)*. National Bureau of Standards, Washington, DC.
- Eugene Garfield. Information, Power, and the *Science Citation Index*. In *Essays of an Information Scientist*, Volume 1, 1962–1973, Institute for Scientific Information.
- Mark Garzone. 1996. *Automated Classification of Citations using Linguistic Semantic Grammars*. M.Sc. Thesis, The University of Western Ontario.
- Mark Garzone and Robert E. Mercer. 2000. Towards an Automated Citation Classifier. In *AI'2000, Proceedings of the 13th Biennial Conference of the CSCSI/SCEIO*, Lecture Notes in Artificial Intelligence, 1822:337–346, H.J. Hamilton (ed.). Springer-Verlag.
- A.G. Gross. 1996. *The Rhetoric of Science*. Harvard University Press.
- A.G. Gross, J.E. Harmon, and M. Reidy. 2002. *Communicating Science: The Scientific Article from the 17th Century to the Present*. Oxford University Press.
- M.A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman Group Limited.
- Ken Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins Publishing Company.
- Alistair Knott. 1996. *A Data-driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, University of Edinburgh.
- B.A. Lipetz. 1965. Problems of Citation Analysis: Critical Review. *American Documentation*, 16:381–390.
- William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281.
- Edward M. Marcotte, Ioannis Xenarios, and David Eisenberg. 2001. Mining Literature for Protein-Protein Interactions. *Bioinformatics*, 17(4):359–363.
- Daniel Marcu. 1997. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph.D. thesis, University of Toronto.
- Robert E. Mercer and Chrysanne Di Marco. 2003. The Importance of Fine-grained Cue Phrases in Scientific Citations. In *AI'2003, Proceedings of the 16th Conference of the CSCSI/SCEIO*, 550–556. Edmonton, Alberta, 11–13 June 2003.
- Robert E. Mercer, Chrysanne Di Marco, and Frederick Kroon. 2004. The Frequency of Hedging Cues in Citation Contexts in Scientific Writing. Submitted to *Conference of the Canadian Society for the Computational Studies of Intelligence (CSCSI 2004)*.
- M.J. Moravcsik and P. Murugesan. 1975. Some Results on the Function and Quality of Citations. *Social Studies of Science*, 5:86–92.
- Greg Myers. 1991. *Writing Biology*. University of Wisconsin Press.
- B.C. Peritz. 1983. A Classification of Citation Roles for the Social Sciences and Related Fields. *Scientometrics*, 5:303–312.
- H. Small. 1978. Cited Documents as Concept Symbols. *Social Studies of Science*, 8(3):327–340.
- I. Spiegel-Rösing. 1977. Science Studies: Bibliometric and Content Analysis. *Social Studies of Science*, 7:97–113.
- James Thomas, David Milward, Christos Ouzounis, Stephen Pulman, and Mark Carroll. 2000. Automatic Extraction of Protein Interactions from Scientific Abstracts. In *Proceedings of the 5th Pacific Symposium on Biocomputing (PSB 2000)*, 538–549.
- M. Weinstock. 1971. Citation Indexes. In *Encyclopaedia of Library and Information Science*, 5:16–40. Marcel Dekker, New York.
- W. John Wilbur. 2002. A Thematic Analysis of the AIDS Literature. In *Proceedings of the 7th Pacific Symposium on Biocomputing (PSB 2004)*, 386–397.
- W.J. Wilbur and L. Coffee. 1994. The Effectiveness of Document Neighboring in Search Enhancement. *Information Processing Management*, 30:253–266.