Argumentative Feedback: A Linguistically-motivated Term Expansion for Information Retrieval

Patrick Ruch, Imad Tbahriti, Julien Gobeill
Medical Informatics Service
University of Geneva
24 Micheli du Crest
1201 Geneva
Switzerland
{patrick.ruch,julien.gobeill,imad.tbahriti}@hcuge.ch

Alan R. Aronson
Lister Hill Center
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894
USA
alan@nlm.nih.gov

Abstract

We report on the development of a new automatic feedback model to improve information retrieval in digital libraries. Our hypothesis is that some particular sentences, selected based on argumentative criteria, can be more useful than others to perform well-known feedback information retrieval tasks. The argumentative model we explore is based on four disjunct classes, which has been very regularly observed in scientific reports: PURPOSE, METHODS, RESULTS, CONCLUSION. To test this hypothesis, we use the Rocchio algorithm as baseline. While Rocchio selects the features to be added to the original query based on statistical evidence, we propose to base our feature selection also on argumentative criteria. Thus, we restrict the expansion on features appearing only in sentences classified into one of our argumentative categories. Our results, obtained on the OHSUMED collection, show a significant improvement when expansion is based on PURPOSE (mean average precision = +23%) and CONCLUSION (mean average precision = +41%) contents rather than on other argumentative contents. These results suggest that argumentation is an important linguistic dimension that could benefit information retrieval.

1 Introduction

Information retrieval (IR) is a challenging endeavor due to problems caused by the underlying expressiveness of all natural languages. One of these problems, synonymy, is that authors and users frequently employ different words or expressions to refer to the same meaning (*accident* may be expressed as *event*, *incident*, *problem*, *difficulty*, *unfortunate situation*, *the subject of your last letter*, *what happened last week*, etc.) (Furnas et al., 1987). Another problem is ambiguity, where a specific term may have several (and sometimes contradictory) meanings and interpretations (e.g., the word *horse* as in *Trojan horse*, *light horse*, *to work like a horse*, *horse about*). In order to obtain better meaning-based matches between queries and documents, various propositions have been suggested, usually without giving any consideration to the underlying domain.

During our participation in different international evaluation campaigns such as the TREC Genomics track (Hersh, 2005), the BioCreative initiative (Hirschman et al., 2005), as well as in our attempts to deliver advanced search tools for biologists (Ruch, 2006) and healthcare providers (Ruch, 2002) (Ruch, 2004), we were more concerned with domain-specific information retrieval in which systems must return a ranked list of MEDLINE records in response to an expert's information request. This involved a set of available queries describing typical search interests, in which gene, protein names, and diseases were often essential for an effective retrieval. Biomedical publications however tend to generate new information very rapidly and also use a wide variation in terminology, thus leading to the current situation whereby a large number of names, symbols and synonyms are used to denote the same concepts. Current solutions to these issues can be classified into domain-specific strategies, such as thesaurus-based expansion, and domain-independent strategies, such as blind-feedback. By proposing to explore a third type of approach, which attempts to take advantage of argumentative specificities of scientific reports, our study initiates a new research direction for natural language processing applied to information retrieval.

The rest of this paper is organized as follows. Section 2 presents some related work in information retrieval and in argumentative parsing, while Section 3 depicts the main characteristics of our test collection and the metrics used in our experiments. Section 4 details the strategy

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 675–682,
Sydney, July 2006.

used to develop our improved feedback method. Section 5 reports on results obtained by varying our model and Section 6 contains conclusions on our experiments.

2 Related works

Our basic experimental hypothesis is that some particular sentences, selected based on argumentative categories, can be more useful than others to support well-known feedback information retrieval tasks. It means that selecting sentences based on argumentative categories can help focusing on content-bearing sections of scientific articles.

2.1 Argumentation

Originally inspired by corpus linguistics studies (Orasan, 2001), which suggests that scientific reports (in chemistry, linguistics, computer sciences, medicine...) exhibit a very regular logical distribution -confirmed by studies conducted on biomedical corpora (Swales, 1990) and by ANSI/ISO professional standards - the argumentative model we experiment is based on four disjunct classes: PURPOSE, METHODS, RESULTS, CONCLUSION.

Argumentation belongs to discourse analysis[1], with fairly complex computational models such as the implementation of the rhetorical structure theory proposed by (Marcu, 1997), which proposes dozens of rhetorical classes. More recent advances were applied to document summarization. Of particular interest for our approach, Teufel and Moens (Teufel and Moens, 1999) propose using a list of manually crafted triggers (using both words and expressions such as *we argued*, *in this article*, *the paper is an attempt to*, *we aim at*, etc.) to automatically structure scientific articles into a lighter model, with only seven categories: BACKGROUND, TOPIC, RELATED WORK, PURPOSE, METHOD, RESULT, and CONCLUSION.

More recently and for knowledge discovery in molecular biology, more elaborated models were proposed by (Mizuta and Collier, 2004) (Mizuta et al., 2005) and by (Lisacek et al., 2005) for novelty-detection. (McKnight and Srinivasan, 2003) propose a model very similar to our four-class model but is inspired by clinical trials. Preliminary applications were proposed for bibliometrics and related-article search (Tbahriti et al., 2004) (Tbahriti et al., 2005), information extraction and passage retrieval (Ruch et al., 2005b). In these studies, sentences were selected as the basic classification unit in order to avoid as far as possible co-reference issues (Hirst, 1981), which hinder readibity of automatically generated and extracted sentences.

2.2 Query expansion

Various query expansion techniques have been suggested to provide a better match between user information needs and documents, and to increase retrieval effectiveness. The general principle is to expand the query using words or phrases having a similar or related meaning to those appearing in the original request. Various empirical studies based on different IR models or collections have shown that this type of search strategy should usually be effective in enhancing retrieval performance. Scheme propositions such as this should consider the various relationships between words as well as term selection mechanisms and term weighting schemes (Robertson, 1990). The specific answers found to these questions may vary; thus a variety of query expansion approaches were suggested (Efthimiadis, 1996).

In a first attempt to find related search terms, we might ask the user to select additional terms to be included in a new query, e.g. (Velez et al., 1997). This could be handled interactively through displaying a ranked list of retrieved items returned by the first query. Voorhees (Voorhees, 1994) proposed basing a scheme based on the WordNet thesaurus. The author demonstrated that terms having a lexical-semantic relation with the original query words (extracted from a synonym relationship) provided very little improvement (around 1% when compared to the original unexpanded query).

As a second strategy for expanding the original query, Rocchio (Rocchio, 1971) proposed accounting for the relevance or irrelevance of top-ranked documents, according to the user's manual input. In this case, a new query was automatically built in the form of a linear combination of the term included in the previous query and terms automatically extracted from both the relevant documents (with a positive weight) and non-relevant items (with a negative weight). Empirical studies (e.g., (Salton and Buckley, 1990)) demonstrated that such an approach is usually quite effective, and could

[1]After Aristotle, discourses structured following an appropriate argumentative distribution belong to logics, while ill-defined ones belong to rhetorics.

be used more than once per query (Aalbersberg, 1992). Buckley et al. (Singhal et al., 1996b) suggested that we could assume, without even looking at them or asking the user, that the top k ranked documents are relevant. Denoted the pseudo-relevance feedback or blind-query expansion approach, this approach is usually effective, at least when handling relatively large text collections.

As a third source, we might use large text corpora to derive various term-term relationships, using statistically or information-based measures (Jones, 1971), (Manning and Schütze, 2000). For example, (Qiu and Frei, 1993) suggested that terms to be added to a new query could be extracted from a similarity thesaurus automatically built through calculating co-occurrence frequencies in the search collection. The underlying effect was to add idiosyncratic terms to the underlying document collection, related to the query terms by language use. When using such query expansion approaches, we can assume that the new terms are more appropriate for the retrieval of pertinent items than are lexically or semantically related terms provided by a general thesaurus or dictionary. To complement this global document analysis, (Croft, 1998) suggested that text passages (with a text window size of between 100 to 300 words) be taken into account. This local document analysis seemed to be more effective than a global term relationship generation.

As a forth source of additional terms, we might account for specific user information needs and/or the underlying domain. In this vein, (Liu and Chu, 2005) suggested that terms related to the user's intention or scenario might be included. In the medical domain, it was observed that users looking for information usually have an underlying scenario in mind (or a typical medical task). Knowing that the number of scenarios for a user is rather limited (e.g., *diagnosis*, *treatment*, *etiology*), the authors suggested automatically building a semantic network based on a domain-specific thesaurus (using the Unified Medical Language System (UMLS) in this case). The effectiveness of this strategy would of course depend on the quality and completeness of domain-specific knowledge sources. Using the well-known term frequency (tf)/inverse document frequency (idf) retrieval model, the domain-specific query-expansion scheme suggested by Liu and Chu (2005) produces better retrieval performance than a scheme based on statistics (MAP: 0.408 without query expansion, 0.433 using statistical methods and 0.452 with domain-specific approaches).

In these different query expansion approaches, various underlying parameters must be specified, and generally there is no single theory able to help us find the most appropriate values. Recent empirical studies conducted in the context of the TREC Genomics track, using the OHSUGEN collection (Hersh, 2005), show that neither blind expansion (Rocchio), nor domain-specific query expansion (thesaurus-based Gene and Protein expansion) seem appropriate to improve retrieval effectiveness (Aronson et al., 2006) (Abdou et al., 2006).

3 Data and metrics

To test our hypothesis, we used the OHSUMED collection (Hersh et al., 1994), originally developed for the TREC topic detection track, which is the most popular information retrieval collection for evaluating information search in library corpora. Alternative collections (cf. (Savoy, 2005)), such as the French Amaryllis collection, are usually smaller and/or not appropriate to evaluate our argumentative classifier, which can only process English documents. Other MEDLINE collections, which can be regarded as similar in size or larger, such as the TREC Genomics 2004 and 2005 collections are unfortunately more domain-specific since information requests in these collection are usually targeting a particular gene or gene product.

Among the 348,566 MEDLINE citations of the OHSUMED collection, we use the 233,455 records provided with an abstract. An example of a MEDLINE citation is given in Table 1: only Title, Abstract, MeSH and Chemical (RN) fields of MEDLINE records were used for indexing. Out of the 105 queries of the OHSUMED collection, only 101 queries have at least one positive relevance judgement, therefore we used only this subset for our experiments. The subset has been randomly split into a training set (75 queries), which is used to select the different parameters of our retrieval model, and a test set (26 queries), used for our final evaluation.

As usual in information retrieval evaluations, the mean average precision, which computes the precision of the engine at different levels (0%, 10%, 20%... 100%) of recall, will be used in our experiments. The precision of the top returned

Title: Computerized extraction of coded findings from free-text radiologic reports. Work in progress.
Abstract: A computerized data acquisition tool, the special purpose radiology understanding system (SPRUS), has been implemented as a module in the Health Evaluation through Logical Processing Hospital Information System. This tool uses semantic information from a diagnostic expert system to parse free-text radiology reports and to extract and encode both the findings and the radiologists' interpretations. These coded findings and interpretations are then stored in a clinical data base. The system recognizes both radiologic findings and diagnostic interpretations. Initial tests showed a true-positive rate of 87% for radiographic findings and a bad data rate of 5%. Diagnostic interpretations are recognized at a rate of 95% with a bad data rate of 6%. Testing suggests that these rates can be improved through enhancements to the system's thesaurus and the computerized medical knowledge that drives it. This system holds promise as a tool to obtain coded radiologic data for research, medical audit, and patient care.
MeSH Terms: *Artificial Intelligence*; Decision Support Techniques; Diagnosis, Computer-Assisted; Documentation; Expert Systems; Hospital Information Systems*; Human; Natural Language Processing*; Online Systems; Radiology Information Systems*.*

Table 1: MEDLINE records with, title, abstract and keyword fields as provided by MEDLINE librarians: major concepts are marked with *; Subheadings and checktags are removed.

document, which is obviously of major importance is also provided together with the total number of relevant retrieved documents for each evaluated run.

4 Methods

To test our experimental hypothesis, we use the Rocchio algorithm as baseline. In addition, we also provide the score obtained by the engine before the feedback step. This measure is necessary to verify that feedback is useful for querying the OHSUMED collection and to establish a strong baseline. While Rocchio selects the features to be added to the original queries based on pure statistical analysis, we propose to base our feature expansion also on argumentative criteria. That is, we overweight features appearing in sentences classified in a particular argumentative category by the argumentative categorizer.

4.1 Retrieval engine and indexing units

The easyIR system is a standard vector-space engine (Ruch, 2004), which computes state-of-the-art *tf.idf* and probabilistic weighting schema. All experiments were conducted with pivoted normalization (Singhal et al., 1996a), which has recently shown some effectiveness on MEDLINE corpora (Aronson et al., 2006). Query and document weighings are provided in Equation (1): the dtu formula is applied to the documents, while the dtn formula is applied to the query; t the number of indexing terms, df_j the number of documents in which the term t_j; pivot and slope are constants (fixed at pivot = 0.14, slope = 146).

$$\begin{array}{l} \text{dtu: } w_{ij} = \frac{(Ln(Ln(tf_{ij})+1)+1)\cdot idf_j}{(1-slope)\cdot pivot+slope\cdot nt_i} \\ \text{dtn: } w_{ij} = idf_j \cdot (Ln(Ln(tf_{if})+1)+1) \end{array} \quad (1)$$

As already observed in several linguistically-motivated studies (Hull, 1996), we observe that common stemming methods do not perform well on MEDLINE collections (Abdou et al., 2006), therefore indexing units are stored in the inverted file using a simple S-stemmer (Harman, 1991), which basically handles most frequent plural forms and exceptions of the English language such as *-ies*, *-es* and *-s* and exclude endings such as *-aies*, *-eies*, *-ss*, etc. This simple normalization procedure performs better than others and better than no stemming. We also use a slightly modified standard stopword list of 544 items, where strings such as *a*, which stands for *alpha* in chemistry and is relevant in biomedical expressions such as *vitamin a*.

4.2 Argumentative categorizer

The argumentative classifier ranks and categorizes abstract sentences as to their argumentative classes. To implement our argumentative categorizer, we rely on four binary Bayesian classifiers, which use lexical features, and a Markov model, which models the logical distribution of the argumentative classes in MEDLINE abstracts. A comprehensive description of the classifier with feature selection and comparative evaluation can be found in (Ruch et al., 2005a)

To train the classifier, we obtained 19,555 explicitly structured abstracts from MEDLINE. A

Abstract: PURPOSE: The overall prognosis for patients with congestive heart failure is poor. Defining specific populations that might demonstrate improved survival has been difficult [...] PATIENTS AND METHODS: We identified 11 patients with severe congestive heart failure (average ejection fraction 21.9 +/- 4.23% (+/- SD) who developed spontaneous, marked improvement over a period of follow-up lasting 4.25 +/- 1.49 years [...] RESULTS: During the follow-up period, the average ejection fraction improved in 11 patients from 21.9 +/- 4.23% to 56.64 +/- 10.22%. Late follow-up indicates an average ejection fraction of 52.6 +/- 8.55% for the group [...] CONCLUSIONS: We conclude that selected patients with severe congestive heart failure can markedly improve their left ventricular function in association with complete resolution of heart failure [...]

Table 2: MEDLINE records with explicit argumentative markers: PURPOSE, (PATIENTS and) METHODS, RESULTS and CONCLUSION.

	Bayesian classifier			
	PURP.	METH.	RESU.	CONC.
PURP.	80.65 %	0 %	3.23 %	16 %
METH.	8 %	78 %	8 %	6 %
RESU.	18.58 %	5.31 %	52.21 %	23.89 %
CONC.	18.18 %	0 %	2.27 %	79.55 %

	Bayesian classifier with Markov model			
	PURP.	METH.	RESU.	CONC.
PURP.	93.35 %	0 %	3.23 %	3 %
METH.	3 %	78 %	8 %	6 %
RESU.	12.73 %	2.07 %	57.15 %	10.01 %
CONC.	2.27 %	0 %	2.27 %	95.45 %

Table 3: Confusion matrix for argumentative classification. The harmonic means between recall and precision score (or F-score) is in the range of 85% for the combined system.

conjunctive query was used to combine the following four strings: *PURPOSE:*, *METHODS:*, *RESULTS:*, *CONCLUSION:*. From the original set, we retained 12,000 abstracts used for training our categorizer, and 1,200 were used for fine-tuning and evaluating the categorizer, following removal of explicit argumentative markers. An example of an abstract, structured with explicit argumentative labels, is given in Table 2. The per-class performance of the categorizer is given by a contingency matrix in Table 3.

4.3 Rocchio feedback

Various general query expansion approaches have been suggested, and in this paper we compared ours with that of Rocchio. In this latter case, the system was allowed to add m terms extracted from the k best-ranked abstracts from the original query. Each new query was derived by applying the following formula (Equation 2): $Q' = \alpha \cdot Q + (\beta/k) \cdot \sum kj = 1w_{ij}$ (2), in which Q' denotes the new query built from the previous query Q, and w_{ij} denotes the indexing term weight attached to the term t_j in the document D_i. By direct use of the training data, we determine the optimal values of our model: m = 10, k = 15. In our experiments, we fixed α = 2.0, β = 0.75. Without feedback the mean average precision of the evaluation run is 0.3066, the Rocchio feedback (mean average precision = 0.353) represents an improvement of about 15% (cf. Table 5), which is statistically[2] significant ($p < 0.05$).

4.4 Argumentative selection for feedback

To apply our argumentation-driven feedback strategy, we first have to classify the top-ranked abstracts into our four argumentative moves: PURPOSE, METHODS, RESULTS, and CONCLUSION. For the argumentative feedback, different m and k values are recomputed on the training queries, depending on the argumentative category we want to over-weight. The basic segment is the sentence; therefore the abstract is split into a set of sentences before being processed by the argumentative classifier. The sentence splitter simply applies as set of regular expressions to locate sentence boundaries. The precision of this simple sentence splitter equals 97% on MEDLINE abstracts. In this setting only one argumentative category is attributed to each sentence, which makes the decision model binary.

Table 4 shows the output of the argumentative classifier when applied to an abstract. To determine the respective value of each argumentative contents for feedback, the argumentative categorizer parses each top-ranked abstract. These abstracts are then used to generate four groups of sentences. Each group corresponds to a unique argumentative class. Each argumentative index contains sentences classified in one of four argumentative classes. Because argumen-

[2]Tests are computed using a non-parametric signed test, cf. (Zobel, 1998) for more details.

CONCLUSION (00160116) The highly favorable pathologic stage (RI-RII, 58%) and the fact that the majority of patients were alive and disease-free suggested a more favorable prognosis for this type of renal cell carcinoma.
METHODS (00160119) Tumors were classified according to well-established histologic criteria to determine stage of disease; the system proposed by Robson was used.
METHODS (00162303) Of 250 renal cell carcinomas analyzed, 36 were classified as chromophobe renal cell carcinoma, representing 14% of the group studied.
PURPOSE (00156456) In this study, we analyzed 250 renal cell carcinomas to a) determine frequency of CCRC at our Hospital and b) analyze clinical and pathologic features of CCRCs.
PURPOSE (00167817) Chromophobe renal cell carcinoma (CCRC) comprises 5% of neoplasms of renal tubular epithelium. CCRC may have a slightly better prognosis than clear cell carcinoma, but outcome data are limited.
RESULTS (00155338) Robson staging was possible in all cases, and 10 patients were stage 1) 11 stage II; 10 stage III, and five stage IV.

Table 4: Output of the argumentative categorizer when applied to an argumentatively structured abstract after removal of explicit markers. For each row, the attributed class is followed by the score for the class, followed by the extracted text segment. The reader can compare this categorization with argumentative labels as provided in the original abstract (PMID 12404725).

tative classes are equally distributed in MEDLINE abstracts, each index contains approximately a quarter of the top-ranked abstracts collection.

5 Results and Discussion

All results are computed using the treceval program, using the top 1000 retrieved documents for each evaluation query. We mainly evaluate the impact of varying the feedback category on the retrieval effectiveness, so we separately expand our queries based a single category. Query expansion based on RESULTS or METHODS sentences does not result in any improvement. On the contrary, expansion based on PURPOSE sentences improve the Rocchio baseline by + 23%, which is again significant ($p < 0.05$). But the main improvement is observed when CONCLUSION sentences are used to generate the expansion, with a remarkable gain of 41% when compared to Rocchio. We also observe in Table 5 that other measures (top precision) and number of relevant retrieved articles do confirm this trend.

For the PURPOSE category, the optimal k parameter, computed on the test queries was 11. For the CONCLUSION category, the optimal k parameter, computed on the test queries was 10. The difference between the m values between Rocchio feedback and the argumentative feedback, respectively 15 vs. 11 and 10 for Rocchio, PURPOSE, CONCLUSION sentences can be explained by the fact that less textual material is available when a particular class of sentences is selected; therefore the number of words that should be added to the original query is more targeted.

No feeback		
Relevant retrieved	Top precision	Mean average precision
1020	0.3871	0.3066
Rocchio feedback		
Relevant retrieved	Top precision	Mean average precision
1112	0.4020	0.353
Argumentative feedback: PURPOSE		
Relevant retrieved	Top precision	Mean average precision
1136	0.485	0.4353
Argumentative feedback: CONCLUSION		
Relevant retrieved	Top precision	Mean average precision
1143	0.550	0.4999

Table 5: Results without feedback, with Rocchio and with argumentative feedback applied on PURPOSE and CONCLUSION sentences. The number of relevant document for all queries is 1178.

From a more general perspective, the importance of CONCLUSION and PURPOSE sentences is consistent with other studies, which aimed at selecting highly content bearing sentences for information extraction (Ruch et al., 2005b). This result is also consistent with the state-of-the-art in automatic summarization, which tends to prefer sentences appearing at the beginning or at the end of documents to generate summaries.

6 Conclusion

We have reported on the evaluation of a new linguistically-motivated feedback strategy, which selects highly-content bearing features for expansion based on argumentative criteria. Our simple model is based on four classes, which have been reported very stable in scientific reports of all kinds. Our results suggest that argumentation-driven expansion can improve retrieval effectiveness of search engines by more than 40%. The proposed methods open new research directions and are generally promising for natural language processing applied to information retrieval, whose positive impact is still to be confirmed (Strzalkowski et al., 1998). Finally, the proposed methods are important from a theoretical perspective, if we consider that it initiates a *genre-specific* paradigm as opposed to the usual information retrieval typology, which distinguishes between domain-specific and domain-independent approaches.

Acknowledgements

The first author was supported by a visiting faculty grant (ORAU) at the Lister Hill Center of the National Library of Medicine in 2005. We would like to thank Dina Demner-Fushman, Susanne M. Humphrey, Jimmy Lin, Hongfang Liu, Miguel E. Ruiz, Lawrence H. Smith, Lorraine K. Tanabe, W. John Wilbur for the fruitful discussions we had during our weekly TREC meetings at the NLM. The study has also been partially supported by the Swiss National Foundation (Grant 3200-065228).

References

I Aalbersberg. 1992. Incremental Relevance Feedback. In *SIGIR*, pages 11–22.

S Abdou, P Ruch, and J Savoy. 2006. General vs. Specific Blind Query Expansion for Biomedical Searches. In *TREC 2005*.

A Aronson, D Demner-Fushman, S Humphrey, J Lin, H Liu, P Ruch, M Ruiz, L Smith, L Tanabe, and J Wilbur. 2006. Fusion of Knowledge-intensive and Statistical Approaches for Retrieving and Annotating Textual Genomics Documents. In *TREC 2005*.

J Xu B Croft. 1998. Corpus-based stemming using cooccurrence of word variants. *ACM-Transactions on Information Systems*, 16(1):61–81.

E Efthimiadis. 1996. Query expansion. *Annual Review of Information Science and Technology*, 31.

G Furnas, T Landauer, L Gomez, and S Dumais. 1987. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11).

D Harman. 1991. How effective is suffixing ? *JASIS*, 42 (1):7–15.

W Hersh, C Buckley, T Leone, and D Hickam. 1994. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *SIGIR*, pages 192–201.

W Hersh. 2005. Report on the trec 2004 genomics track. pages 21–24.

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. 2005. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 (suppl. 1).

G Hirst. 1981. *Anaphora in Natural Language Understanding: A Survey*. Lecture Notes in Computer Science 119 - Springer.

D Hull. 1996. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society of Information Science*, 47(1):70–84.

K Sparck Jones. 1971. *Automatic Keyword Classification for Information Retrieval*. Butterworths.

F Lisacek, C Chichester, A Kaplan, and Sandor. 2005. Discovering Paradigm Shift Patterns in Biomedical Abstracts: Application to Neurodegenerative Diseases. In *Proceedings of the First International Symposium on Semantic Mining in Biomedicine (SMBM)*, pages 212–217. Morgan Kaufmann.

Z Liu and W Chu. 2005. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *ACM-SAC Information Access and Retrieval Track*, pages 1076–1083.

C Manning and H Schütze. 2000. *Foundations of Statistical Natural Language Processing*. MIT Press.

D Marcu. 1997. The Rhetorical Parsing of Natural Language Texts. pages 96–103.

L McKnight and P Srinivasan. 2003. Categorization of sentence types in medical abstracts. *AMIA Annu Symp Proc.*, pages 440–444.

Y Mizuta and N Collier. 2004. Zone identification in biology articles as a basis for information extraction. *Proceedings of the joint NLPBA/BioNLP Workshop on Natural Language for Biomedical Applications*, pages 119–125.

Y Mizuta, A Korhonen, T Mullen, and N Collier. 2005. Zone Analysis in Biology Articles as a Basis for Information Extraction. *International Journal of Medical Informatics*, to appear.

C Orasan. 2001. Patterns in Scientific Abstracts. In *Proceedings of Corpus Linguistics*, pages 433–445.

Y Qiu and H Frei. 1993. Concept based query expansion. *ACM-SIGIR*, pages 160–69.

S Robertson. 1990. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364.

J Rocchio. 1971. *Relevance feedback in information retrieval in The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall.

P Ruch, R Baud, C Chichester, A Geissbühler, F Lisacek, J Marty, D Rebholz-Schuhmann, I Tbahriti, and AL Veuthey. 2005a. Extracting Key Sentences with Latent Argumentative Structuring. In *Medical Informatica Europe (MIE)*, pages 835–40.

P Ruch, L Perret, and J Savoy. 2005b. Features Combination for Extracting Gene Functions from MEDLINE. In *European Colloquium on Information Retrieval (ECIR)*, pages 112–126.

P Ruch. 2002. Using contextual spelling correction to improve retrieval effectiveness in degraded text collections. *COLING 2002*.

P Ruch. 2004. Query translation by text categorization. *COLING 2004*.

P Ruch. 2006. Automatic Assignment of Biomedical Categories: Toward a Generic Approach. *Bioinformatics*, 6.

G Salton and C Buckley. 1990. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4).

J Savoy. 2005. Bibliographic database access using free-text and controlled vocabulary: An evaluation. *Information Processing and Management*, 41(4):873–890.

A Singhal, C Buckley, and M Mitra. 1996a. Pivoted document length normalization. *ACM-SIGIR*, pages 21–29.

C Buckley A Singhal, M Mitra, and G Salton. 1996b. New retrieval approaches using smart. *In Proceedings of TREC-4*.

T Strzalkowski, G Stein, G Bowden Wise, J Perez Carballo, P Tapanainen, T Jarvinen, A Voutilainen, and J Karlgren. 1998. Natural language information retrieval: TREC-7 report. In *Text REtrieval Conference*, pages 164–173.

J Swales. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.

I Tbahriti, C Chichester, F Lisacek, and P Ruch. 2004. Using Argumention to Retrieve Articles with Similar Citations from MEDLINE. *Proceedings of the joint NLPBA/BioNLP Workshop on Natural Language for Biomedical Applications*.

I Tbahriti, C Chichester, F Lisacek, and P Ruch. 2005. Using Argumentation to Retrieve Articles with Similar Citations: an Inquiry into Improving Related Articles Search in the MEDLINE Digital Library. *International Journal of Medical Informatics*, to appear.

S Teufel and M Moens. 1999. Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting. *Advances in Automatic Text Summarization, MIT Press*, pages 155–171.

B Velez, R Weiss, M Sheldon, and D Gifford. 1997. Fast and effective query refinement. In *ACM SIGIR*, pages 6–15.

E Voorhees. 1994. Query expansion using lexical-semantic relations. In *ACM SIGIR*, pages 61–69.

J Zobel. 1998. How reliable are large-scale information retrieval experiments? *ACM-SIGIR*, pages 307–314.