A Study of Structured Clinical Abstracts and the Semantic Classification of Sentences

Grace Y. Chung and Enrico Coiera
Centre for Health Informatics
University of New South Wales
Sydney NSW 2052 Australia
{graceyc, e.coiera}@unsw.edu.au

Abstract

This paper describes experiments in classifying sentences of medical abstracts into a number of semantic classes given by section headings in structured abstracts. Using conditional random fields, we obtain F-scores ranging from 0.72 to 0.97. By using a small set of sentences that appear under the PARTICPANTS heading, we demonstrate that it is possible to recognize sentences that describe population characteristics of a study. We present a detailed study of the structure of abstracts of randomized clinical trials, and examine how sentences labeled under PARTICIPANTS could be used to summarize the population group.

1 Introduction

Medical practitioners are increasingly applying evidence-based medicine (EBM) to support decision-making in patient treatments. The aim of EBM (Sackett, 1998) is to provide improved care leading to better outcomes through locating evidence for a clinical problem, evaluating the quality of the evidence, and then applying to a current problem at hand. However, the adoption of EBM is hampered by the overwhelming amount of information available, and insufficient time and skills on the clinician's part to locate and synthesize the best evidence in the scientific literature.

MEDLINE abstracts about randomized clinical trials (RCTs) play a critical role in providing the best evidence for the latest interventions for any given conditions. The MEDLINE database now has 16 million bibliographic entries, many of them include the abstract and more than 3 million of these were published in the last 5 years (Hunter, 2006).

To alleviate the information overload, some resources such as the Cochrane Collaboration (Cochrane, 2007), Evidence-Based Medicine (EBM, 2007), the ACP Journal Club (ACP, 2007) and BMJ Clinical Evidence (BMJCE, 2007), employ human experts to summarize knowledge within RCTs through extensive searches and critical assessments.

In (Sim, 2000), RCT information is entered into electronic knowledge bases or *"trial banks"*, easing the task for systematic reviewing and critical appraisal. This project requires manual entry of descriptions about the design and execution (subjects, recruitment, treatment assignment, follow-up), and hence, only small numbers of RCTs have been archived thus far.

The goal of our research is to use natural language processing to extract the most important pieces of information from RCTs for the purpose of automatic summarization, tailored towards the medical practitioner's clinical question at hand. Ultimately, it is our vision that data mined from full text articles of RCTs not only aid clinicians' assessments but researchers who are conducting meta-analyses.

In this paper, we examine the use of section headings that are frequently given in abstracts of medical journal articles. These section headings are topic-independent. Effectively they define the discourse structure for the abstract, and provide semantic labels to the sentences that fall under them.

BioNLP 2007: Biological, translational, and clinical language processing, pages 121–128,
Prague, June 2007.

Other researchers have recognized the potential utility of these heading (McKnight, 2003; Xu, 2006; Lin, 2006). It has also been recognized that scientific abstracts with such labels could be of importance to text summarization, information retrieval and question answering (Lee, 2006; Zweigenbaum, 2003). We share similar goals to previous research; the section headings of these structured medical abstracts can be used as training data for building labelers that can tag unstructured abstracts with discourse structure. But also, there is a large number of heading names. Sentences that occur under these heading names form a labeled training set which could be used to build a classifier that recognizes similar sentences. Ultimately, we would like to build finer-grained classifiers that exploit these semantic labels.

In our work, we seek to demonstrate that information about patient characteristics can now be extracted from structured and unstructured abstracts. We are motivated by the fact that patient characteristics is one of the fundamental factors most pertinent to evaluation of relevance to a clinical question. The total number of subjects in a trial reflects on the quality of the RCT, and additional factors such as age, gender and other co-existing conditions, will be crucial for assessing whether an RCT is relevant to the medical practitioner's immediate patient.

This paper is organized as follows. In Section 1 we will describe how the RCT abstracts were obtained, and we present a study of the discourse headings that occur in our document corpus. Section 3 will detail our sentence classification experiments. We first explore classification in which the abstracts are labeled under five subheadings, one of which describes the patients or population group. We also perform classification using a combined two-stage scheme, bootstrapping from partially labeled data. Finally in Section 4, we consider how well the PARTICIPANTS labeled sentences capture sentences containing the total number of participants in a trial. In Section 5, we will give a detailed analysis of the labeled sentences.

2 The Data

2.1 Corpus Creation

The current corpus is obtained by a MEDLINE search for RCTs. We did not constrain publications by their date. For the purpose of constraining the size of our corpus in these preliminary experiments, it was our intention to use RCTs pertaining to a fixed set of clinical conditions. Hence, we conducted a MEDLINE search for RCTs with the following keywords: asthma, diabetes, breast cancer, prostate cancer, erectile dysfunction, heart failure, cardiovascular, angina. The resultant corpus contains 7535 abstracts of which 4268 are structured.

2.2 Structure of Medical Abstracts

Structured abstracts were introduced in 1987 (AdHoc, 2005) to help clinical readers to quickly select appropriate articles, and allow more precise information retrieval. However, currently, the majority of medical abstracts remain unstructured. Previous studies have concluded that while many scientific abstracts follow consistent patterns (e.g. Introduction, Problem, Method, Evaluation, Conclusion) many still contain missing sections or have differing structures (Orasan, 2001; Swales, 1990; Meyer, 1990). Journals vary widely in their requirements for abstract structures.

We have conducted a study of the structured abstracts in our corpus. Of 4268 structured abstracts, we have found a total of 238 unique section headings. The most common ones are shown in Table 1. To investigate the numbers of variations in the abstract structure, we first manually map headings that

1. RESULTS	6. METHODS / RESULTS
2. METHODS	7. OBJECTIVE
3. CONCLUSION	8. PATIENTS / METHODS
4. BACKGROUND	9. PURPOSE
5. CONCLUSION	10. DESIGN

Table 1: The most common headings in RCT abstracts.

Class	Example Heading Names
Aim	AIM, AIMS, AIM OF THE STUDY..
Setting	SETTING, SETTINGS, STUDY SETTING..
Participants	PARTICIPANTS, PATIENTS, SUBJECTS..
Setting/ Subjects	PARTICIPANTS AND SETTINGS, SETTING/PATIENTS..

Table 2: Examples of manual mappings for heading names into equivalence classes.

Structure of Abstracts	% of Corpus
BACKGROUND, METHOD, RESULT, CONCLUSION	16%
AIM, METHOD, RESULT, CONCLUSION	14%
AIM, PATIENT AND METHOD, RESULT, CONCLUSION	8.5%
BACKGROUND, AIM, METHOD, RESULT, CONCLUSION	7.6%
BACKGROUND, METHOD AND RESULTS, CONCLUSION	6.6%
AIM, PARTICIPANTS, DESIGN, MEASUREMENTS, RESULT, CONCLUSION	<1%
CONTEXT, DESIGN, SETTING, PARTICIPANTS, OUTCOME MEASURES, RESULT, CONCLUSION	<1%
AIM, DESIGN AND SETTING, PARTICIPANTS, INTERVENTION MEASUREMENTS AND MAIN RESULTS, CONCLUSION	<1%

Table 3: Examples of the patterns that occur in the section headings of structured RCT abstracts.

are essentially semantically equivalent to the same classes, resulting in 106 classes. Examples of these mappings are shown in Table 2. After the class mappings are applied, it turns out that there are still 400 different patterns in the combinations of section headings in these medical abstracts, with over 90% of these variations occurring less than 10 times. The most common section heading patterns are shown in Table 3. Some of the less common ones are also shown.

In studying the structure of these medical abstracts, we find that the variation in structural ordering is large, and many of the heading names are unique, chosen at the discretion of the paper author. Some of the most frequent heading names are also compound headings such as: METHODS/RESULTS, RESULTS/CONCLUSION, PATIENTS/RESULTS, SUBJECTS AND SETTINGS.

3 Sentence Classification Experiments

3.1 Extracting Participant Sentences

In this work, we seek to build a classifier using training data from the semantic labels already provided by structured abstracts. It is our intention ultimately to label both structured and unstructured abstracts with the semantic labels that are of interest for the purposes of information extraction and answering specific questions regarding the trial. In our approach, we identify in our structured abstracts the ones with section headings about patient characteristics. These are collapsed under one semantic class and used as training data for a classifier.

From our 4268 structured abstracts, all the heading names are examined and are re-mapped by hand to one of five heading names: AIM, METHOD, PARTICIPANTS, RESULTS, CONCLUSION. Most heading names can be mapped to these general headings but the subset containing compound headings such as METHOD/RESULT are discarded.

All the abstracts are segmented into sentences and tokenized via Metamap (Aronson, 2001). Some abstracts are discarded due to sentence segmentation errors. The remainder (3657 abstracts) forms the corpus that we will work with here. These abstracts are randomly divided into a training set and an initial test set, and for purposes of our experiments, they are further subdivided into abstracts with the PARTICIPANTS label and those without. The exact size of our data sets are given in Table 4.

Although abstracts in Train Set A are generally structured as (AIM, METHOD, RESULTS, CONCLUSION), they contain sentences pertaining to patient or population group largely in the METHOD section. In the following, we will explore three ways for labeling sentences in the abstract including labeling for sentences that describe the population group. The first employs a 5-class classifier, the second uses a two-stage approach and the third employs an approach which uses partially labeled data.

3.2 Using Labeled Data Only

Using only abstracts from Train Set B, all sentences are mapped into one of 5 classes: AIM, PARTICIPANTS, METHOD, RESULTS, CONCLUSION.

Data Set	Number of Abstracts	Number of Sentences
Total in Corpus	3657	45k
Total Train Set	3439	42k
Train Set A (no PARTICIPANTS)	2643	32k
Train Set B (w/ PARTICIPANTS)	796	10k
Test Set (w/ PARTICIPANTS)	62	878

Table 4: Sizes of data sets.

	Recall	Precision	F-score
CRF	Accuracy = 84.4%		
Aim	0.98	0.91	0.95
Method	0.52	0.73	0.61
Participants	0.79	0.73	0.76
Results	0.95	0.87	0.91
Conclusion	0.91	0.97	0.94
SVM	Accuracy = 80.2%		
Aim	0.87	0.91	0.90
Method	0.64	0.68	0.67
Participants	0.73	0.70	0.72
Results	0.89	0.84	0.86
Conclusion	0.80	0.88	0.83

Table 5: Classification of sentences in RCT abstracts into 5 semantic classes using CRFs and SVMs. The recall, precision and F-score are reported on our unseen test set.

The PARTICIPANTS class subsume all headings that include mention of population characteristics. These include compound headings such as: SETTING/POPULATION, PATIENTS/DESIGN. Sentences associated with these compound headings often include long sentences that describe the participant group as well as a second aspect of the study such as setting or design.

We build a 5-class classifier using linear-chain conditional random fields (CRFs).[1] CRFs (Sutton, 2006) are undirected graphical models that are discriminatively trained to maximize the conditional probability of a set of output variables given a set of input variables. We simply use bag-of-words as features because past studies (McKnight, 2003), using n-gram-based features did not improve accuracies.[2]

As a baseline comparison, we have performed classification using a Support Vector Machine (SVM) classifier (Burges, 1998; Witten, 2005), with a radial basis functions (RBF) kernel. To help model the sequential ordering, a normalized integer for the sentence number in the abstract is included as a feature.

Experimental results are shown in Table 5. CRFs clearly outperform SVMs in this classification task. This may in part be attributable to the explicit sequential modeling in the CRFs compared with SVMs. While our training set (796 abstracts in Train set B) is substantially smaller than that reported in previous studies (McKnight, 2003; Lin, 2006; Xu, 2006), the F-score for AIM, RESULTS, CONCLUSION are comparable to previous results. By far the largest sources of classification error are the confusions between METHOD and PARTICIPANTS class. In training we have included into the PARTICIPANTS class all sentences that come under compound headings, and therefore the PARTICIPANTS section can often encompass several sentences that contain detailed information regarding the intervention, and the type of study, as exemplified below.

> Doppler echocardiography was performed in 21 GH deficient patients after 4 months placebo and 4 months GH therapy, in a double blind cross-over study. In an open design study, 13 patients were reinvestigated following 16 months and 9 patients following 38 months of GH therapy. Twenty-one age and sex-matched normal control subjects were also investigated.

Nonetheless, information about the patient population is embedded within these sentences.

3.3 Using a Two-Stage Method

An alternative approach is to adopt a two-stage hierarchical strategy. First we build a classifier which performs a 4-way classification based on the labels AIM, METHOD, RESULTS, CONCLUSION, and a second stage binary classifier tags all the METHOD sentences into either METHOD or PARTICIPANTS. There are two distinct advantages to this approach. (1) In our 5-class classifier, it is clear that METHOD and PARTICIPANTS are confusable and a dedicated classifier to perform this subtask may be more effective. (2) The corpus of abstracts with only the 4 classes labeled is much larger (3439 abstracts), and hence the resultant classifier is likely to be trained more robustly. Our first stage classifier is a CRF tagger. It is trained on the combined training sets A and B, whereby all sentences in the structured abstracts are mapped to the 4-class labels. The second stage binary classifier is an SVM classifier. The SVM classifier has been augmented with additional features of the semantic labels tagged via Metamap tagger. It is trained on the subset of Train Set A (3499 sentences) that is labeled as either METHOD or PARTICIPANTS.

Classification results for the unseen test set are reported in Table 6. The 4-class classifier yields F-scores between 0.92 and 0.96. We report results for

[1] We used the SimpleTagger command line interface of the Mallet software package (McCallum, 2002).

[2] In other experiments, attempts to use stemming and removal of stop words also did not improve performance.

(1) 4-class	Accuracy = 92.7%		
	Recall	Precision	F-score
Aim	0.98	0.94	0.96
Method	0.89	0.95	0.92
Results	0.95	0.89	0.92
Conclusion	0.91	0.97	0.94
(2) 2-class	Accuracy = 80.1%		
Method	0.73	0.83	0.78
Participants	0.87	0.78	0.81
(3) 5-class	Accuracy = 86.0%		
Aim	0.96	0.92	0.96
Method	0.66	0.79	0.71
Participants	0.77	0.72	0.75
Results	0.94	0.89	0.92
Conclusion	0.91	0.97	0.94

Table 6: (1) Classification using CRFs into 4 major semantic classes with combined Train Set A and B as training data. (2) Binary SVM classification of a subset of test set sentences. (3) Classification into 5 classes as described in Section 3.3. All results (recall, precision and F-score) are reported on the unseen test set.

the binary SVM classifier on the subset of test set sentences (253 sentences) that are either METHOD or PARTICIPANTS in Table 6.

The two stage method here has yielded some gains in performance for each class except for PARTICIPANTS. The gains are likely to have been due to increased training data particularly for the classes, AIM, RESULTS and CONCLUSION.

3.4 Augmenting with Partially Labeled Data

We investigate a second method for leveraging the data available in Train Set A. We hypothesize that many sentences within the METHOD section of Train Set A do in fact describe patient information and could be used as training data. We propose a bootstrapping method whereby some of the sentences in Train Set A are tagged by a binary SVM classifier and used as training data in the 5-class CRF classifier. The following describes each step:

1. A binary SVM classifier is trained on the subset of sentences in Train Set B labeled with METHOD and PARTICIPANTS.
2. The trained SVM classifier is used to label all the sentences in Train Set A that are originally labeled with the METHOD class.
3. All the sentences in Train Set A are now labeled in terms of the 5 classes, and a score is available from the SVM output is associated with those sentences labeled as either METHOD or PARTICIPANTS. The abstracts that contain sentences scoring above a pre-determined threshold score are then pooled with sentences in Train Set B into a single training corpus. We tuned the threshold value by testing on a development set held out from Train Set B. As a result, 1217 sentences from Train Set A is combined with Train Set B.
4. The final training corpus is used to train a CRF tagger to label sentences into one of 5 classes.

	Recall	Precision	F-score
5-class	Accuracy = 87.6%		
Aim	0.99	0.95	0.97
Method	0.67	0.77	0.72
Participants	0.90	0.77	0.83
Results	0.91	0.92	0.92
Conclusion	0.90	0.97	0.93

Table 7: Classification into 5 classes as described in Section 3.4. All results (recall, precision and F-score) are reported on the unseen test set.

The results of classification on the unseen test set are reported in Table 7. Overall accuracy for classification improves to 87.6% primarily because there is a marked improvement is observed for the F-scores of the PARTICIPANTS class. Our best results here are comparable to those previously reported on similar tasks on the class, AIM, RESULTS and CONCLUSION (Xu, 2006; Lin, 2006). The F-score for METHOD is lower because introducing a PARTICIPANTS label has increased confusability.

4 Extraction of Number of Patients

We have demonstrated that for a structured abstract it is possible to predict sentences that are associated with population characteristics. However, our ultimate objective is to extract these kinds of sentences from unstructured abstracts, and even to extract more fine-grained information. In this section, we will examine whether labeling sentences into one of 5 classes can aid us in the extraction of the total number of patients from an RCT.

	Abstracts w/ Total Subjects	% tagged as PARTICIPANTS
Structured	46	87%
Unstructured	103	72%

Table 8: Extraction of the total number of subjects in a trial in a human annotated test set, as described in Section 4.2

4.1 Annotation

In a concurrent annotation effort to label RCT abstracts, human annotators manually tagged a separate test set of 204 abstracts with the total number of participants in each study. Of the 204 abstracts, 148 are unstructured and 56 are structured. None of these 204 abstracts are part of the training set, described in this paper.

4.2 Experiments

The abstracts from this annotated test set are processed by the classifier described in Section 3.4. For all the abstracts which mention the total number of participants in the RCT, we compute the frequency for which this is included in the sentences labeled as PARTICIPANTS. Results are depicted in Table 8.

Upon subsequent examination of the test set, it is found that only 82% (46/56) of the structured abstracts and 70% (103/148) of unstructured abstracts contain information about total number of participants in the trial. As seen in Table 8, in 87% of the 46 structured abstracts, and in 72% of the 103 unstructured abstracts, the total number of participants are mentioned in the labeled PARTICIPANTS sentences. The extraction of the total number of participants is significantly worse in unstructured abstracts which do not adhere to the strict discourse structures given by the headings of structured abstracts. In 13% (13/103) of the unstructured abstracts, the total number of participants appears in the first sentence, which is usually tagged as the AIM. It is evident that in the absence of structure, patient information can occur in any sentence in the abstract, or for that matter, it may appear only in the body of the paper. Our method of training first on structured abstracts may be a strong limitation to extraction of information from unstructured abstracts.

Even for the structured abstracts in the test set, 9% (4/46) of the set of abstracts containing population number actually mention the number in the AIM or RESULTS section, rather than the METHOD or PARTICIPANTS. Only 12 abstracts contain explicit headings referring to participants, where the total number of subjects in the trial is mentioned under the corresponding heading.

In this task, we only consider that total number of subjects enrolled in a study, and have yet to account for additional population numbers such as the drop out rate, the follow-up rate, or the number of subjects in each arm of a study. These are often reported in an abstract without mentioning the total number of patients to begin with. The classifier will tag sentences that describe these as PARTICIPANT sentences nonetheless.

5 Analysis and Discussion

We will further analyze the potential for using sentences tagged as PARTICIPANTS as summaries of population characteristics for a trial. Table 9 gives some examples of sentences tagged by the classifier.

Sentences that appear under PARTICIPANTS in structured abstracts are often concise descriptions of the population group with details about age, gender, and conditions, as seen in Example 1. Otherwise, they can also be extensive descriptions, providing selection criteria and some detail about method, as in Example 2.

Examples 3 and 4 show sentences from the test set of Section 4. Example 3 has been labeled as a PARTICIPANTS sentence by the classifier. It describes patient characteristics, giving the population number for each arm of the trial but does not reveal the total number of subjects. Example 3 appears under the heading METHODS AND RESULTS in the original abstract. Example 4 is from an unstructured abstract, where information about the intervention and population and study design are interleaved in the same sentences but tagged by the classifier as PARTICIPANTS. Many sentences tagged as PARTICIPANTS also do not give explicit information about population numbers but only provide descriptors for patient characteristics.

It is also plausible that our task has been made more challenging compared with previous reported studies because our corpus has not been filtered for publication date. Hence, the numbers of publica-

1. Male smokers aged 50–69 years who had angina pectoris in the Rose chest pain questionnaire at baseline (n = 1795). *PMID: 9659191*
2. The study included 809 patients under 70 years of age with stable angina pectoris. The mean age of the patients was 59 +/- 7 years and 31% were women. Exclusion criteria were myocardial infarction within the previous 3 years and contraindications to beta-blockers and calcium antagonists. The patients were followed between 6 and 75 months (median 3.4 years and a total of 2887 patient years). *PMID: 8682134*
3. Subjects with Canadian Cardiovascular Society (CCS) class 3/4 angina and reversible perfusion defects were randomized to SCS (34) or PMR (34). *PMID: 16554313*
4. Sixty healthy women, half of whom had been using OCs for at least the previous 6 months, participated in the study. Approximately two thirds were smokers and were randomized to be tested after either a 12 hr nicotine deprivation or administration of nicotine gum. One third were nonsmokers. *PMID: 11495215*

Table 9: Examples of sentences labeled under PARTICIPANTS class, forming summaries of the population characteristics of a trial. Examples 1 and 2 are typical sentences under the PARTICIPANTS heading in the train set. Examples 3 and 4 are from the annotated test set. See Section 5 for more detailed explanation.

tions and structural characteristics of our abstracts may be broader than previous reports which filter for abstracts to a narrow time frame (Xu, 2006).

6 Related Work

In recent years, there has been a growth in research in information extraction and NLP in the medical domain particularly in the RCT literature. This is due in part to the emergence of lexical and semantic resources such as the Unified Medical Language System (UMLS) (Lindberg, 1993), and software such as MetaMap (Aronson, 2001), which transforms text into UMLS concepts, and SemRep (Rindflesch, 2003), which identifies semantic propositions.

There are a number of previous attempts to perform text categorization on sentences in MEDLINE abstracts into generic discourse level section headings. They all share the goal of assigning structure to unstructured abstracts for the purpose of summarization or question answering. All previous attempts have mapped the given headings to four or five generic classes, and performed text categorization on large sets of RCTs without any disease or condition-specific filtering. Studies have shown that results deteriorate when classifying sentences in unstructured abstracts (McKnight, 2003; Lin, 2006). In (McKnight, 2003), McKnight and Srinivisan used an SVM for tagging sentences into 4 classes. Using a corpus of 7k abstracts, they obtain F-scores from 0.82 to 0.89. Later papers in (Xu, 2006; Lin, 2006) have found that Hidden Markov Models (HMMs) based approaches more effectively model the sequential ordering of sentences in abstracts. In (Xu, 2006), several machine learning methods, decision tree, maximum entropy and naive Bayes, are evaluated with an HMM-based algorithm. 3.8k abstracts from 2004 and 2005 were used as training data, and experiments yielded average precision of 0.94 and recall of 0.93.

One driving model for information extraction in RCTs is the PICO framework (Richardson, 1995). This is a task-based model for EBM formulated to assist EBM practitioners to articulate well-formed questions in order to find useful answers in clinical scenarios. PICO elements are Patient/Population, Intervention, Comparison and Outcome. This model has been adopted by researchers (Demner-Fushman, 2005; Niu, 2004) as a guideline for elements that can be automatically extracted from RCTs and patient records. However, doubts have been raised about the utility of PICO as a generic knowledge representation for computational approaches to answering clinical questions (Huang, 2006).

In experiments reported in (Demner-Fushman, 2005), the PICO framework was used as a basis for extracting population, problem, intervention and comparison for the purpose of evaluating relevance of an abstract to a particular clinical question. In this work, the population statements were located via a set of hand-written rules that were based on extracting an actual numeric value for the population.

7 Conclusions

In this study, we investigated the use of conditional random fields for classifying sentences in medical abstracts. Our results particularly in terms of F-scores for generic section headings such as AIM, RE-

SULTS and CONCLUSION were comparable to previous studies, even with smaller training sets. We investigated the use of text classification by leveraging the subset of abstracts with explicitly labeled PARTICIPANTS sentences combining the use of CRFs and SVMs, and exploiting partially labeled data.

One main objective here is to label sentences that describe population characteristics in structured and unstructured abstracts. We found that unstructured abstracts differ substantially from structured ones, and alternative approaches will be necessary for extracting information from unstructured abstracts. Furthermore, critical details that are needed by a physician when evaluating a study such as exclusion criteria, drop out rate, follow up rate, etc, may only be listed in the full text of the study. Future work will address extracting information beyond the abstract.

8 Acknowledgment

The authors would like to acknowledge the anonymous reviewers and the executive committee for their comments and suggestions, and Marianne Byrne, Brenda Anyango Omune and Wei Shin Yu for annotation of the abstracts. This project is funded by the Australian Research Council, grant number DP0666600.

References

ACP Journal Club. Available from: http://www.acpjp.org

Ad Hoc working group for Critical Appraisal of the Medical Literature 1987. A proposal for more informative abstracts of clinical articles. *Annals of Int. Medicine* 106:595–604.

A. R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Ann. Symp. of AMIA* pp 17–21.

Clinical Evidence. BMJ Publishing Group. Available from: http://www.clinicalevidence.com

C. Burges. 1998. A Tutorial on Support Vector Machines for Pattern Recognition *Journal Data Mining and Knowledge Discovery,* 2(2), June.

The Cochrane Collaboration. Available from: http://www.cochrane.org

D. Demner-Fushman and J. Lin. 2005. Knowledge extraction for clinical question answering: Preliminary results. *AAAI Workshop on Question Answering in Restricted Domains*.

Evidence Based Medicine. Available from: http://ebm.bmjjournals.com

X. Huang et al. 2006. Evaluation of PICO as a Knowledge Representation for Clinical Questions. *Ann. Symp. of AMIA* pp359–363.

L. Hunter and K. Bretonnel Cohen. 2006. Biomedical language processing: what's beyond PubMed? *Molecular Cell,* 21:589-594.

J. Lin et al. 2006. Generative Content Models for Structural Analysis of Medical Abstracts. *Workshop on Biomedical Natural Language Processing BioNLP* New York.

D. A. Lindberg et al. 1993. The Unified Medical Language System. *Methods of Information in Medicine,* 32(4):281–291.

A. McCallum. 2002. *MALLET: A Machine Learning for Language Toolkit.* http://mallet.cs.umass.edu.

L. McKnight and P. Srinivasan 2003. Categorization of Sentence Types in Medical Abstracts. *Ann. Symp. of AMIA* pp440–444.

M. Lee et al. 2006. Beyond Information Retrieval–Medical Question Answering. *Ann. Symp. of AMIA.*

Y. Niu and G. Hirst. 2005. Analysis of semantic classes in medical text for question answering. *Workshop on Question Answering in Restricted Domains,* Barcelona.

C. Orasan. 2001. Patterns in Scientific Abstracts. *2001 Corpus Linguistics Conference.*

W. S. Richardson et al. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP J Club,* Nov-Dec;123(3):A12-3.

T. Rindflesch and M. Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *J. of Biomedical Informatics,* 36(6):462–477, Dec.

D. L. Sackett et al.. 1998. *Evidence Based Medicine: How to Practice and Teach EBM.* Churchill Livingstone, Edinburgh.

F. Salager-Meyer. 1990. Discourse Movements in Medical English Abstracts and their linguistic exponents: A genre analysis study. *INTERFACE: J. of Applied Linguistics,* 4(2):107–124.

I. Sim et al. 2000. Electronic Trial Banks: A Complementary Method for Reporting Randomized Trials. *Med Decis Making,* Oct-Dec;20(4):440-50.

C. Sutton and A McCallum. 2006. An introduction to conditional random fields for relational learning. In Lise Getoor and Ben Taskar, editors, *Introduction to Statistical Relational Learning.* MIT Press. To appear.

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

I. H. Witten and E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques,* 2nd Ed, Morgan Kaufmann, San Francisco.

R. Xu et al. 2006. Combining Text Classification and Hidden Markov Modeling Techniques for Structuring Randomized Clinical Trial Abstracts. *Ann. Symp. of AMIA.*

P. Zweigenbaum. 2003. Question answering in biomedicine. *Workshop on Natural Language Processing for Question Answering,* Budapest.