Positional Language Models for Clinical Information Retrieval

Florian Boudin
DIRO, Université de Montréal
CP. 6128, succ. Centre-ville
H3C 3J7 Montréal, Canada
boudinfl@iro.umontreal.ca

Jian-Yun Nie
DIRO, Université de Montréal
CP. 6128, succ. Centre-ville
H3C 3J7 Montréal, Canada
nie@iro.umontreal.ca

Martin Dawes
Department of Family Medicine
McGill University, 515 Pine Ave
H2W 1S4 Montréal, Canada
martin.dawes@mcgill.ca

Abstract

The PECO framework is a knowledge representation for formulating clinical questions. Queries are decomposed into four aspects, which are Patient-Problem (P), Exposure (E), Comparison (C) and Outcome (O). However, no test collection is available to evaluate such framework in information retrieval. In this work, we first present the construction of a large test collection extracted from systematic literature reviews. We then describe an analysis of the distribution of PECO elements throughout the relevant documents and propose a language modeling approach that uses these distributions as a weighting strategy. In our experiments carried out on a collection of 1.5 million documents and 423 queries, our method was found to lead to an improvement of 28% in MAP and 50% in P@5, as compared to the state-of-the-art method.

1 Introduction

In recent years, the volume of health and biomedical literature available in electronic form has grown exponentially. MEDLINE, the authoritative repository of citations from the medical and bio-medical domain, contains more than 18 million citations. Searching for clinically relevant information within this large amount of data is a difficult task that medical professionals are often unable to complete in a timely manner. A better access to clinical evidence represents a high impact application for physicians.

Evidence-Based Medicine (EBM) is a widely accepted paradigm for medical practice (Sackett et al., 1996). EBM is defined as the conscientious, explicit and judicious use of current best evidence in making decisions about patient care. Practice EBM means integrating individual clinical expertise with the best available external clinical evidence from systematic research. It involves tracking down the best evidence from randomized trials or meta-analyses with which to answer clinical questions. Richardson et al. (1995) identified the following four aspects as the key elements of a well-built clinical question:

- **Patient-problem**: what are the patient characteristics (e.g. age range, gender, etc.)? What is the primary condition or disease?
- **Exposure-intervention**: what is the main intervention (e.g. drug, treatment, duration, etc.)?
- **Comparison**: what is the exposure compared to (e.g. placebo, another drug, etc.)?
- **Outcome**: what are the clinical outcomes (e.g. healing, morbidity, side effects, etc.)?

These elements are known as the PECO elements. Physicians are educated to formulate their clinical questions in respect to this structure. For example, in the following question: "*In patients of all ages with Parkinson's disease, does a Treadmill training compared to no training allows to increase the walking distance*?" one can identify the following elements:

- **P**: Patients of all ages with Parkinson's disease
- **E**: Treadmill training
- **C**: No treadmill training
- **O**: Walking distance

In spite of this well-defined question structure, physicians still use keyword-based queries when they search for clinical evidence. An explanation of

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 108–115,
MIT, Massachusetts, USA, 9-11 October 2010.

that is the almost total absence of PECO search interfaces. PubMed[1], the most used search interface, does not allow users to formulate PECO queries yet. For the previously mentioned clinical question, a physician would use the query "*Treadmill* AND *Parkinson's disease*". There is intuitively much to gain by using a PECO structured query in the retrieval process. This structure specifies the role of each concept in the desired documents, which is a clear advantage over a keyword-based approach. One can for example differentiate two queries in which a disease would be a patient condition or a clinical outcome. This conceptual decomposition of queries is also particularly useful in a sense that it can be used to balance the importance of each element in the search process.

Another important factor that prevented researchers from testing approaches to clinical information retrieval (IR) based on PECO elements is the lack of a test collection, which contains a set of documents, a set of queries and the relevance judgments. The construction of such a test collection is costly in manpower. In this paper, we take advantage of the systematic reviews about clinical questions from Cochrane. Each Cochrane review examines in depth a clinical question and survey all the available relevant publications. The reviews are written for medical professionals. We transformed them into a TREC-like test collection, which contains 423 queries and 8926 relevant documents extracted from MEDLINE. In a second part of this paper, we present a model integrating the PECO framework in a language modeling approach to IR. An intuitive method would try to annotate the concepts in documents into PECO categories. One can then match the PECO elements in the query to the elements detected in documents. However, as previous studies have shown, it is very difficult to automatically annotate accurately PECO elements in documents. To by-pass this issue, we propose an alternative that relies on the observed positional distribution of these elements in documents. We will see that different types of element have different distributions. By weighting words according to their positions, we can indirectly weigh the importance of different types of element in search. As we will show in this paper, this approach turns out to be highly effective.

This paper is organized as follows. We first briefly review the previous work, followed by a description of the test collection we have constructed. Next, we give the details of the method we propose and present our experiments and results. Lastly, we conclude with a discussion and directions for further work.

2 Related work

The need to answer clinical questions related to a patient care using IR systems has been well studied and documented (Hersh et al., 2000; Niu et al., 2003; Pluye et al., 2005). There are a limited but growing number of studies trying to use the PECO elements in the retrieval process. (Demner-Fushman and Lin, 2007) is one of the few such studies, in which a series of knowledge extractors is used to detect PECO elements in documents. These elements are later used to re-rank a list of retrieved citations from PubMed. Results reported indicate that their method can bring relevant citations into higher-ranking positions, and from these abstracts generate responses that answer clinicians' questions. This study demonstrates the value of the PECO framework as a method for structuring clinical questions. However, as the focus has been put on the post-retrieval step (for question-answering), it is not clear whether PECO elements are useful at the retrieval step. Intuitively, the integration of PECO elements in the retrieval process can also lead to higher retrieval effectiveness.

The most obvious scenario for testing this would be to recognize PECO elements in documents prior to indexing. When a PECO-structured query is formulated, it is matched against the PECO elements in the documents (Dawes et al., 2007). Nevertheless, the task of automatically identifying PECO elements is a very difficult one. There are two major reasons for that. First, previous studies have indicated that there is a low to moderate agreement rate among humans for annotating PECO elements. This is due to the lack of standard definition for the element' boundaries (e.g. can be words, phrases or sentences) but also to the existence of several levels of annotation. Indeed, there are a high number

[1] www.pubmed.gov

of possible candidates for each element and one has to choose if it is a main element (i.e. playing a major role in the clinical study) or secondary elements. Second is the lack of sufficient annotated data that can be used to train automatic tagging tools.

Despite all these difficulties, several efficient detection methods have been proposed (Demner-Fushman and Lin, 2007; Chung, 2009). Nearly all of them are however restricted to a coarse-grain annotation level (i.e. tagging entire sentences as describing one element). This kind of coarser-grain identification is more robust and more feasible than the one at concept level, and it could be sufficient in the context of IR. In fact, for IR purposes, what is the most important is to correctly weight the words in documents and queries. From this perspective, an annotation at the sentence level may be sufficient. Notwithstanding, experiments conducted using a collection of documents that were annotated at a sentence-level only showed a small increase in retrieval accuracy (Boudin et al., 2010b) compared to a traditional bag-of-words approach.

More recently, Boudin et al. (2010a) proposed an alternative to the PECO detection issue that relies on assigning different weights to words according to their positions in the document. A location-based weighting strategy is used to emphasize the most informative parts of documents. They show that a large improvement in retrieval effectiveness can be obtained this way and indicate that the weights learned automatically are correlated to the observed distribution of PECO elements in documents. In this work, we propose to go one step further in this direction by analyzing the distribution of PECO elements in a large number of documents and define the positional probabilities of PECO elements accordingly. These probabilities will be integrated in the document language model.

3 Construction of the test collection

Despite the increasing use of search engines by medical professionals, there is no standard test collection for evaluating clinical IR. Constructing such a resource from scratch would require considerable time and money. One way to overcome this obstacle is to use already available systematic reviews. Systematic reviews try to identify, appraise, select and synthesize all high quality research evidence relevant to a clinical question. The best-known source of systematic reviews in the healthcare domain is the Cochrane collaboration[2]. It consists of a group of over 15,000 specialists who systematically identify and review randomized trials of the effects of treatments. In particular, a review contains a reference section, listing all the relevant studies to the clinical question. These references can be considered as relevant documents. In our work, we propose to use these reviews as a way to semi-automatically build a test collection. As the reviews are made by specialists in the area independently from our study, we can avoid bias in our test collection.

We gathered a subset of Cochrane systematic reviews and asked a group of annotators, one professor and four Master students in family medicine, to create PECO-structured queries corresponding to the clinical questions. As clinical questions answered in these reviews cover various aspects of one topic, multiple variants of precise PECO queries were generated for each review. Moreover, in order to be able to compare a PECO-based search strategy to a real world scenario, this group have also provided the keyword-based queries that they would have used to search with PubMed. Below is an example of queries generated from the systematic review about "*Aspirin with or without an antiemetic for acute migraine headaches in adults*":

Keyword-based query

[aspirin and migraine]

PECO-structured queries

1. [adults 18 years or more with migraine]^{P}
   [aspirin alone]^{E}
   [placebo]^{C}
   [pain free]^{O}
2. [adults 18 years or more with migraine]^{P}
   [aspirin plus an antiemetic]^{E}
   [placebo]^{C}
   [pain free]^{O}
3. [adults 18 years or more with migraine]^{P}
   [aspirin plus metoclopramide]^{E}
   [active comparator]^{C}
   [use of rescue medication]^{O}

[2] www.cochrane.org

All the citations included in the "References" section of the systematic review were extracted and selected as relevant documents. These citations were manually mapped to PubMed unique identifiers (PMID). This is a long process that was undertaken by two different workers to minimize the number of errors. At this step, only articles published in journals referenced in PubMed are considered (e.g. conference proceedings are not included).

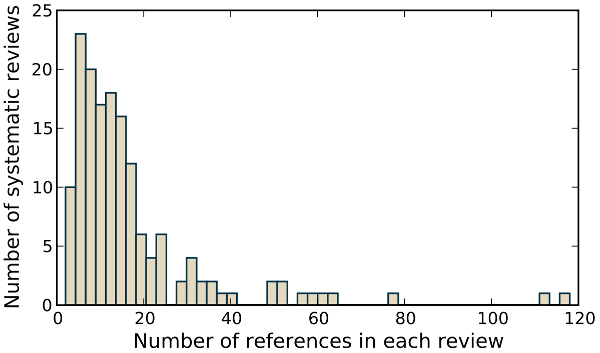

Figure 1: Histogram of the number of queries versus the number of relevant documents.

We selected in sequential order from the set of new systematic reviews[3] and processed 156 Cochrane reviews. There was no restriction about the topics covered or the number of included references. The resulting test collection is composed of 423 queries and 8926 relevant citations (2596 different citations). This number reduces to 8138 citations once we remove the citations without any text in the abstract (i.e. certain citations, especially old ones, only contain a title). Figure 1 shows the statistics derived from the number of relevant documents by query. In this test collection, the average number of documents per query is approximately 19 while the average length of a document is 246 words.

4 Distribution of PECO elements

The observation that PECO elements are not evenly distributed throughout the documents is not new. In fact, most existing tagging methods used location-based features. This information turns out to be very useful because of the standard structure of medical citations. Actually, many scientific journals explicitly recommend authors to write their abstracts in compliance to the ordered rhetorical structure: Introduction, Methods, Results and Discussion. These rhetorical categories are highly correlated to the distributions of PECO elements, as some elements are more likely to occur in certain categories (e.g. clinical outcomes are more likely to appear in the conclusion). The position is thus a strong indicator of whether a text segment contains a PECO element or not.

To the best of our knowledge, the first analysis of the distribution of PECO elements in documents was described in(Boudin et al., 2010a). A small collection of manually annotated abstracts was used to compute the probability that a PECO element occurs in a specific part of the documents. This study is however limited by the small number of annotated documents (approximately 50 citations) and the moderate agreement rate among human annotators. Here we propose to use our test collection to compute more reliable statistics.

The idea is to use the pairs of PECO-structured query and relevant document, assuming that if a document is relevant then it should contain the same elements as the query. Of course, this is obviously not always the case. Errors can be introduced by synonyms or homonyms and relevant documents may not contain all of the elements described in the query. But, with more than 8100 documents, it is quite safe to say that this method produce fairly reliable results. Moreover, a filtering process is applied to queries removing all non-informative words (e.g. stopwords, numbers, etc.) from being counted.

There are several ways to look at the distribution of PECO elements in documents. One can use the rhetorical structure of abstracts to do that. However, the high granularity level of such analysis would make it less precise for IR purposes. Furthermore, most of the citations available in PubMed are devoid of explicitly marked sections. It is possible to automatically detect these sections but only with a non-negligible error rate (McKnight and Srinivasan, 2003). In our study, we chose to use a fixed number of partitions by dividing documents into parts of equal length. This choice is motivated by its repeatability and ease to implement, but also for comparison with previous studies.

We divided each relevant document into 10 parts of equal length on a word level (from P1 to P10). We

[3] `http://mrw.interscience.wiley.com/cochrane/cochrane_clsysrev_new_fs.html`

computed statistics on the number of query words that occur in each of these parts. For each PECO element, the distribution of query words among the parts of the documents is not uniform (Figure 2). We observe distinctive distributions, especially for Patient-Problem and Exposure elements, indicating that first and last parts of the documents have higher chance to contain these elements. This gives us a clear and robust indication on which specific parts should be enhanced when searching for a given element. Our proposed model will exploit the typical distributions of PECO elements in documents.

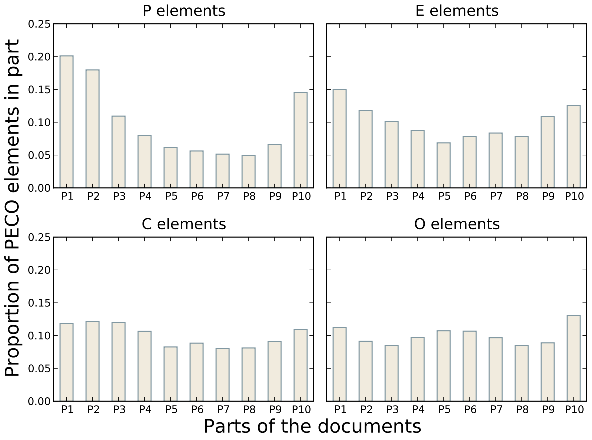

Figure 2: Distribution of each PECO element throughout the different parts of the documents.

5 Retrieval Method

In this work, we use the language modeling approach to information retrieval. This approach assumes that queries and documents are generated from some probability distribution of text (Ponte and Croft, 1998). Under this assumption, ranking a document D as relevant to a query Q is seen as estimating $P(Q|D)$, the probability that Q was generated by the same distribution as D. A typical way to score a document D as relevant to a query Q is to compute the Kullback-Leibler divergence between their respective language models:

$$\text{score}(Q, D) = \sum_{w \in Q} P(w|Q) \cdot \log P(w|D) \quad (1)$$

Under the traditional bag-of-words assumption, i.e. assuming that there is no need to model term dependence, a simple estimate for $P(w|Q)$ can be obtained by computing Maximum Likelihood Estimation (MLE). It is calculated as the number of times the word w appears in the query Q, divided by its length:

$$P(w|Q) = \frac{\text{count}(w, Q)}{|Q|}$$

A similar method is employed for estimating $P(w|D)$. Bayesian smoothing using Dirichlet priors is however applied to the maximum likelihood estimator to compensate for data sparseness (i.e. smoothing probabilities to remove zero estimates). Given μ the prior parameter and $\mathcal{C}$ the collection of documents, $P(w|D)$ is computed as:

$$P(w|D) = \frac{\text{count}(w, D) + \mu \cdot P(w|\mathcal{C})}{|D| + \mu}$$

5.1 Model definition

In our model, we propose to use the distribution of PECO elements observed in documents to emphasize the most informative parts of the documents. The idea is to get rid of the problem of precisely detecting PECO elements by using a positional language model. To integrate position, we estimate a series of probabilities that constraints the word counts to a specific part of the documents instead of the entire document. Each document D is ranked by a weighted linear interpolation. Given a document D divided in 10 parts $p \in [P1, P2 \cdots P10]$, $P(w|D)$ in equation 1 is redefined as:

$$\begin{aligned} P'(w|D) = \alpha \cdot P(w|D) + \beta \cdot P_{\text{title}}(w|D) \\ + \gamma \cdot \sum_{p_i \in D} \sigma_e \cdot P_{p_i}(w|D) \end{aligned} \quad (2)$$

where the σ_e weights for each type of element e are empirically fixed to the values of the distribution of PECO elements observed in documents. We then redefine the scoring function to integrate the PECO query formulation. The idea is to use the PECO structure as a way to balance the importance of each element in the retrieval step. The final scoring function is defined as:

$$\text{score}_{\text{final}}(Q, D) = \sum_{e \in \text{PECO}} \delta_e \cdot \text{score}(Q_e, D)$$

In our model, there are a total of 7 weighting parameters, 4 corresponding to the PECO elements in queries (δ_P, δ_E, δ_C and δ_O) and 3 for the document language models (α, β and γ). These parameters will be determined by cross-validation.

6 Results

In this section, we first describe the details of our experimental protocol. Then, we present the results obtained by our model on the constructed test collection.

6.1 Experimental settings

As a collection of documents, we gathered 1.5 millions of citations from PubMed. We used the following constraints: citations with an abstract, human subjects, and belonging to one of the following publication types: randomized control trials, reviews, clinical trials, letters, editorials and meta-analyses. The set of queries and relevance judgments described in Section 3 is used to evaluate our model. Relevant documents were, if not already included, added to the collection. Because each query is generated from a systematic literature review completed at a time t, we placed an additional restriction on the publication date of the retrieved documents: only documents published before time t are considered. Before indexing, each citation is pre-processed to extract its title and abstract text and then converted into a TREC-like document format. Abstracts are divided into 10 parts of equal length (the ones containing less than 10 words are discarded). The following fields are marked in each document: title, P1, P2 $\cdots$ P10. The following evaluation measures are used:

- Precision at rank n (P@n): precision computed on the n topmost retrieved documents.
- Mean Average Precision (MAP): average of precision measures computed at the point of each relevant document in the ranked list.
- Number of relevant documents retrieved

All retrieval tasks are performed using an "out-of-the-shelf" version of the Lemur toolkit[4]. We use the embedded tokenization algorithm along with the standard Porter stemmer. The number of retrieved documents is set to 1000 and the Dirichlet prior smoothing parameter to $\mu = 2000$. In all our experiments, we use the KL divergence scoring function (equation 1) as baseline. Statistical significance is computed using the well-known Student's t-test. To determine reasonable weights and avoid overtuning the parameters, we use a 10-fold cross-validation optimizing the MAP values.

6.2 Experiments

We first investigated the impact of using PECO-structured queries on the retrieval performance. As far as we know, no quantitative evaluation of the increase or decrease of performance in comparison with a keyword-based search strategy has been reported. Schardt et al. (2007) presented a comparison between PubMed and a PECO search interface but failed to demonstrate any significant difference between the two search protocols. The larger number of words in PECO-structured queries, on average 18.8 words per query compared to 4.3 words for keyword queries, should capture more aspects of the information need. But, it may also be a disadvantage due to the fact that more noise can be brought in, causing query-drift issues.

We propose two baselines using the keyword-based queries. The first baseline (named Baseline-1) uses keyword queries with the traditional language modeling approach. This is one of the state-of-the-art approaches in current IR research. This retrieval model considers each word in a query as an equal, independent source of information. In the second baseline (named Baseline-2), we consider multiword phrases. In our test collection, queries are often composed of multiword phrases such as "*low back pain*" or "*early pregnancy*". It is clear that finding the exact phrase "*heart failure*" is a much stronger indicator of relevance than just finding "*heart*" and "*failure*" scattered within a document. The Indri operator `#1` is used to perform phrase-based retrieval. Phrases are already indicated in queries by the conjunction and (e.g. *vaccine and hepatitis B*). A simple regular expression is used to recognize the phrases.

Results are presented in Table 1. As expected, phrase-based retrieval leads to some increase in retrieval precision (P@5). However, the number of

[4] www.lemurproject.org

relevant documents retrieved is decreased. This is due to the fact that we use exact phrase matching that can reduce query coverage. One solution would be to use unordered window features (Indri operator `#uwn`) that would require words to be close together but not necessarily in an exact sequence order (Metzler and Croft, 2005).

The PECO queries use PECO-structured queries as a bag of words. We observe that PECO queries do not enhance the average precision but increase the P@5 significantly. The number of relevant documents retrieved is also larger. These results indicate that formulating clinical queries according to the PECO framework enhance the retrieval effectiveness.

Model	MAP	P@5	#rel. ret.
Baseline-1	**0.129**	0.151	5369
Baseline-2	0.128	0.161*	4645
PECO-queries	0.126	**0.172***	**5433**

Table 1: Comparing the performance measures of keyword-based and PECO-structured queries in terms of MAP, precision at 5 and number of relevant documents retrieved (#rel. ret.). ($*$: t.test $<$ 0.05)

In a second series of experiments, we evaluated the model we proposed in Section 5 . We compared two variants of our model. The first variant (named Model-1) uses a global σ_e distribution fixed according to the average distribution of all PECO elements (i.e. the observed probability that a PECO element occurs in a document' part, no matter which element it is). The second variant (named Model-2) uses a differentiated σ_e distribution for each type of PECO element. The idea is to see if, given the fact that PECO elements have different distributions in documents, using an adapted weight distribution for each element can improve the retrieval effectiveness.

Previous studies have shown that assigning a different weight to each PECO element in the query leads to better results (Demner-Fushman and Lin, 2007; Boudin et al., 2010a). In order to compare our model with a similar method, we defined another baseline (named Baseline-3) by fixing the parameters $\beta = 0$ and $\gamma = 0$ in equation 2. We performed a grid search (from 0 to 1 by step of 0.1) to find the optimal δ weights. Regarding the last three parameters in our full models, namely α, β and γ, we conducted a second grid search to find their optimal values. Performance measures obtained in 10-fold cross-validation (optimizing the MAP measure) by these models are presented in Table 2.

A significant improvement is obtained by the Baseline-3 over the keyword-based approach (Baseline-2). The PECO decomposition of queries is particularly useful to balance the importance of each element in the scoring function. We observe a large improvement in retrieval effectiveness for both models over the two baselines. This strongly indicates that a weighting scheme based on the word position in documents is effective. These results support our assumption that the distribution of PECO elements in documents can be used to weight words in the document language model.

However, we do not observe meaningful differences between Model-1 and Model-2. This tend to suggest that a global distribution is likely more robust for IR purposes than separate distributions for each type of element. Another possible reason is that our direct mapping from positional distribution to probabilities may not be the most appropriate. One may think about using a different transformation, or performing some smoothing. We will leave this for our future work.

7 Conclusion

This paper first presented the construction of a test collection for evaluating clinical information retrieval. From a set of systematic reviews, a group of annotators were asked to generate structured clinical queries and collect relevance judgments. The resulting test collection is composed of 423 queries and 8926 relevant documents. This test collection provides a basis for researchers to experiment with PECO-structured queries in clinical IR. The test collection introduced in this paper, along with the manual given to the group of annotators, will be available for download[5].

In a second step, this paper addressed the problem of using the PECO framework in clinical IR. A straightforward idea is to identify PECO elements in documents and use the elements in the retrieval process. However, this approach does not work well be-

[5] http://www-etud.iro.umontreal.ca/~boudinfl/pecodr/

Model	MAP	% rel.	P@5	% rel.	#rel. ret.
Baseline-2	0.128	-	0.161	-	4645
Baseline-3	0.144	+12.5%*	0.196	+21.7%†	5780
Model-1	0.164	+28.1%†	0.241	+49.7%†	5768
Model-2	0.163	+27.3%†	0.240	+49.1%†	5770

Table 2: 10-fold cross validation scores for the Baseline-2, Baseline-3 and the two variants of our proposed model (Model-1 and Model-2). Relative increase over the Baseline-2 is given, #rel. ret. is the number of relevant documents retrieved. ($\dagger$: t.test < 0.01, $*$: t.test < 0.05)

cause of the difficulty to automatically detect these elements. Instead, we proposed a less demanding approach that uses the distribution of PECO elements in documents to re-weight terms in the document model. The observation of variable distributions in our test collection led us to believe that the position information can be used as a robust indicator of the presence of a PECO element. This strategy turns out to be promising. On a data set composed of 1.5 million citations extracted with PubMed, our best model obtains an increase of 28% for MAP and nearly 50% for P@5 over the classical language modeling approach.

In future work, we intend to expand our analysis of the distribution of PECO elements to a larger number of citations. One way to do that would be to automatically extract PubMed citations that contain structural markers associated to PECO categories (Chung, 2009).

References

Florian Boudin, Jian-Yun Nie, and Martin Dawes. 2010a. Clinical Information Retrieval using Document and PICO Structure. In *Proceedings of the HLT-NAACL 2010 conference*, pages 822–830.

Florian Boudin, Lixin Shi, and Jian-Yun Nie. 2010b. Improving Medical Information Retrieval with PICO Element Detection. In *Proceedings of the ECIR 2010 conference*, pages 50–61.

Grace Y. Chung. 2009. Sentence retrieval for abstracts of randomized controlled trials. *BMC Medical Informatics and Decision Making*, 9(1).

Thomas Owens Sheri Keitz Connie Schardt, Martha B Adams and Paul Fontelo. 2007. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics and Decision Making*, 7(1).

Martin Dawes, Pierre Pluye, Laura Shea, Roland Grad, Arlene Greenberg, and Jian-Yun Nie. 2007. The identification of clinically important elements within medical journal abstracts: PatientPopulationProblem, ExposureIntervention, Comparison, Outcome, Duration and Results (PECODR). *Informatics in Primary care*, 15(1):9–16.

D. Demner-Fushman and J. Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103.

William R. Hersh, Katherine Crabtree, David H. Hickam, Lynetta Sacherek, Linda Rose, and Charles P. Friedman. 2000. Factors associated with successful answering of clinical questions using an information retrieval system. *Bulletin of the Medical Library Association*, 88(4):323–331.

Larry McKnight and Padmini Srinivasan. 2003. Categorization of sentence types in medical abstracts. Proceedings of the AMIA annual symposium.

Donald Metzler and W. Bruce Croft. 2005. A Markov random field model for term dependencies. In *Proceedings of the SIGIR conference*, pages 472–479.

Yun Niu, Graeme Hirst, Gregory McArthur, and Patricia Rodriguez-Gianolli. 2003. Answering clinical questions with role identification. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 73–80.

Pierre Pluye, Roland M. Grad, Lynn G. Dunikowski, and Randolph Stephenson. 2005. Impact of clinical information-retrieval technology on physicians: a literature review of quantitative, qualitative and mixed methods studies. *International Journal of Medical Informatics*, 74(9):745–768.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the SIGIR conference*, pages 275–281.

Scott W. Richardson, Mark C. Wilson, Jim Nishikawa, and Robert S. Hayward. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123(3):A12–13.

David L. Sackett, William Rosenberg, J. A. Muir Gray, Brian Haynes, and W. Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't. *British medical journal*, 312:71–72.