# Clinical Information Retrieval using Document and PICO Structure

**Florian Boudin and Jian-Yun Nie**
DIRO, Université de Montréal
CP. 6128, succursale Centre-ville
Montréal, H3C 3J7 Québec, Canada
{boudinfl,nie}@iro.umontreal.ca

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

## Abstract

In evidence-based medicine, clinical questions involve four aspects: Patient/Problem (P), Intervention (I), Comparison (C) and Outcome (O), known as PICO elements. In this paper we present a method that extends the language modeling approach to incorporate both document structure and PICO query formulation. We present an analysis of the distribution of PICO elements in medical abstracts that motivates the use of a location-based weighting strategy. In experiments carried out on a collection of 1.5 million abstracts, the method was found to lead to an improvement of roughly 60% in MAP and 70% in P@10 as compared to state-of-the-art methods.

## 1 Introduction

As the volume of published medical literature continues to grow exponentially, there is more and more research for physicians to assess and evaluate and less time to do so. Evidence-based medicine (EBM) (Sackett et al., 1996) is a widely accepted paradigm in medical practice that relies on evidence from patient-centered clinical research to make decisions. Taking an evidence-based approach to searching means doing a systematic search of all the available literature, individually critically appraising each research study and then applying the findings in clinical practice. However, this is a time consuming activity. One way to facilitate searching for a precise answer is to formulate a well-focused and structured question (Schardt et al., 2007).

Physicians are educated to formulate their clinical questions according to several well defined aspects in EBM: Patient/Problem (P), Intervention (I), Comparison (C) and Outcome (O), which are called PICO elements. In many documents in medical literature (e.g. MEDLINE), one can find the elements of the PICO structure, but rarely explicitly annotated (Dawes et al., 2007). To identify documents corresponding to a patient's state, physicians also construct their queries according to the PICO structure. For example, in the question "*In children with pain and fever how does paracetamol compared with ibuprofen affect levels of pain and fever?*" one can identify the following PICO elements:

Patient/Problem: *children/pain and fever*
Intervention: *paracetamol*
Comparison: *ibuprofen*
Outcome: *levels of pain and fever*

Very little work, if any, has been carried out on the use of these elements in the Information Retrieval (IR) process. There are several reasons for that. It is not easy to identify PICO elements in documents, as well as in the question if these are not explicitly separated in it. Several studies have been performed on identifying PICO elements in abstracts (Demner-Fushman and Lin, 2007; Hansen et al., 2008; Chung, 2009). However, all of them are reporting coarse-grain (sentence-level) tagging methods that have not yet been shown to be sufficient for the purpose of IR. Moreover, there is currently no standard test collection of questions in PICO structure available for evaluation. On the other hand, the most critical aspect in IR is term weighting. One of the purpose of tagging PICO elements is to assign appropriate weights to these elements during the retrieval process. From this perspective, a semantic tagging of PICO elements may be a task that goes well beyond



*Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 822–830,
Los Angeles, California, June 2010. 

that is required for IR. It may be sufficient to have a method that assigns appropriate weights to elements rather than recognizing their semantic roles. In this paper, we will propose an approach to determine term weights according to document structure. This method will be compared to that using tagging of PICO elements.

In this paper, we first report an attempt to manually annotate the PICO elements in documents by physicians and use them as training data to build an automatic tagging tool. It turns out that there is a high disagreement rate between human annotators. The utilization of the automatic tagging tool in an IR experiment shows only a small gain in retrieval effectiveness. We therefore propose an alternative to PICO element detection that uses the structural information of documents. This solution turns out to be robust and effective. The alternative approach is motivated by a strong trend that we observe in the distribution of PICO elements in documents. We then make use of both PICO query and document structure to extend the classical language modeling approach to IR. Specifically, we investigate how each element of a PICO query should be weighted and how a location-based weighting strategy can be used to emphasize the most informative parts (i.e. containing the most PICO elements) of documents.

The paper is organized as follows. We first briefly review the previous work, followed by a description of the method we propose. Next, we present our experiments and results. Lastly, we conclude with a discussion and directions for future work.

## 2 Related work

There have been only a few studies trying to use PICO elements in the retrieval process. (Demner-Fushman and Lin, 2007) is one of the few such studies. The method they describe consists in re-ranking an initial list of retrieved citations. To this end, the relevance of a document is scored by the use of detected PICO elements, among other things. Several other studies aimed to build a Question-Answering system for clinical questions (Demner-Fushman and Lin, 2006; Andrenucci, 2008). But again, the focus has been set on the post-retrieval step, while the document retrieval step only uses a standard approach. In this paper, we argue that IR has much to gain by using PICO elements.

The task of identifying PICO elements has however gain more attention. In their paper, (Demner-Fushman and Lin, 2007) presented a method that uses either manually crafted pattern-matching rules or a combination of basic classifiers to detect PICO elements in medical abstracts. Prior to that, biomedical concepts are labelled by Metamap (Aronson, 2001) while relations between these concepts are extracted with SemRep (Rindflesch and Fiszman, 2003). Recently, supervised classification using Support Vector Machines (SVM) was proposed by (Hansen et al., 2008) to extract the number of trial participants. In a later study, (Chung, 2009) extended this work to other elements using Conditional Random Fields. Although these studies are reporting interesting results, they are limited in several aspects. First, many are restricted to some segments of the medical documents (e.g. Method section) (Chung, 2009), and in most cases, the test collection is very small (a few hundreds abstracts). Second, the precision and granularity of these methods have not yet been shown to be sufficient for the purpose of IR.

The structural information provided by markup languages (e.g. XML) has been successfully used to improve the IR effectiveness (INEX, 2002 2009). For such documents, the structure information can be used to emphasize some particular parts of the document. Thereby, a given word should not have the same importance depending on its position in the document structure.

Taking into account the structure can be done either at the step of querying or at the step of indexing. One way to integrate the structure at querying is to adapt query languages (Fuhr and Großjohann, 2001). These approaches follow the assumption that the user knows where the most relevant information is located. However, (Kamps et al., 2005) showed that it is preferable to use structure as a search hint, and not as a strict search requirement

The second approach consists in integrating the document structure at the indexing step by introducing a structure weighting scheme (Wilkinson, 1994). In such a scheme, the weight assigned to a word is not only based on its frequency but also on its position in the document. The structure of a document can be defined in terms of tags (e.g. title, section),

each of those having a weight chosen either empirically or automatically by the use of optimizing techniques such as genetic algorithms (Trotman, 2005).

## 3 Using PICO elements in retrieval

In this section, we present an experiment on the manual annotation of PICO elements. We then describe an approach to detect these elements in documents and give some results on the use of these tagged elements in the retrieval process.

### 3.1 Manual annotation of PICO elements

We asked medical professionals to manually annotate the PICO elements in a small collection of abstracts from PubMed[1]. The instructions given to the annotators were fairly simple. They were asked to precisely annotate all PICO elements in abstracts with no restriction about the size of the elements (i.e. they could be words, phrases or sentences). More than 50 abstracts were manually annotated this way by at least two different annotators. Two annotations by two annotators are considered to agree if they share some words (i.e. they overlap). We computed the well known Cohen's kappa measure as well as an ad-hoc measure called loose. The latter uses PICO elements as units and estimates the proportion of elements that have been annotated by both raters.

| Measure | P-element | I/C-element | O-element |
|---|---|---|---|
| kappa | 0.687 | 0.539 | 0.523 |
| loose | 0.363 | 0.136 | 0.140 |

Table 1: Agreement measures computed for each element. Cohen's kappa and loose agreement are presented.

We can observe that there is a very low agreement rate between human annotators. The loose measure indicates that less than 15% of the I, C and O elements have been marked by both annotators. This fact shows that such human annotations can be hardly used to develop an automatic tagging tool for PICO elements, which requires consistent training data. We therefore try to develop a coarser-grained tagging method.

[1] www.pubmed.gov, PubMed is a service of the US National Library of Medicine that includes over 19 million citations from MEDLINE and other life science journals.

### 3.2 Automatic detection of PICO elements

Similarly to previous work, we propose a sentence-level detection method. The identification of PICO elements can be seen as a classification task. Even for a coarser-grain classification task, we are still lack of annotated data. One solution is to use the structural information embedded in some medical abstracts for which the authors have clearly stated distinctive sentence headings. Some recent abstracts in PubMed do contain explicit headings such as "PATIENTS", "SAMPLE" or "OUTCOMES", that can be used to locate sentences corresponding to PICO elements. Using that information, we extracted three sets of abstracts: Patient/Problem (14 279 abstracts), Intervention/Comparison (9 095) and Outcome (2 394).

Tagging each document goes through a three steps process. First, the document is segmented into plain sentences. Then each sentence is converted into a feature vector using statistical (e.g. position, length) and knowledge-based features (e.g. MeSH semantic type). Knowledge-based features were derived either from manually crafted cue-words/verbs lists or semantic types within the MeSH ontology[2]. Finally, each vector is submitted to multiple classifiers, one for each element, allowing to label the corresponding sentence. We use several algorithms implemented in the Weka toolkit[3]: decision trees, SVM, multi-layer perceptron and Naive Bayes. Combining multiple classifiers using a weighted linear combination of their prediction scores achieves the best results with a f-measure score of 86.3% for P, 67% for I/C and 56.6% for O in 10-fold cross-validation.

### 3.3 Use of detected elements in IR

We use language modeling approach to IR in this work. The idea is that a document is a good match to a query if its language model is likely to generate the query (Ponte and Croft, 1998). It is one of the state-of-the-art approaches in current IR research. Most language modeling work in IR use unigram language models –also called bags-of-words models– assuming that there is no structure in queries or documents. A typical way to score a document d as relevant to a query q is to use the *Kullback-Leibler*

[2] www.nlm.nih.gov/mesh/

[3] www.cs.waikato.ac.nz/ml/index.html

divergence between their respective LMs:

$$\begin{aligned} score(q, d) &= \sum_{w \in q} P(w \mid M_q) \cdot \log P(w \mid M_d) \quad (1) \\ &\propto -KL(M_q \parallel M_d) \end{aligned}$$

where $M_q$ is the LM of the query and $M_d$ the LM of the document. $P(w \mid M_\rho)$ estimates the probability of the word $w$ given the language model $M_\rho$. The most direct way to estimate these models is to use Maximum Likelihood estimation over the words:

$$P(w \mid M_\rho) = \frac{count(w, \rho)}{\mid \rho \mid}$$

where $\rho$ is the observed document, $count(w, \rho)$ the number of times the word $w$ occurs in $\rho$ and $\mid \rho \mid$ the length of the document. Bayesian smoothing using *Dirichlet* priors is then applied to the maximum likelihood estimator to compensate for data sparseness.

We propose an approach that extend the basic LM approach to take into consideration the PICO element annotation. We assume that each element in the document has a different importance weight. Four more LMs are created, one for each elements. Given $\omega_e$ the weight of the PICO element $e$, $P(w \mid M_d)$ in equation 1 is re-defined as:

$$P_1(w \mid M_d) \propto P(w \mid M_d) + \sum_{e \in [P,I,C,O]} \omega_e \cdot P(w \mid M_e)$$

The right hand of the above equation is not a probability function. We could use a normalization to transform it. However, for the purpose of document ranking, this will not make any difference. Therefore, we will keep the un-normalized value.

We performed an extensive series of experiments using this model on the test collection described in Section 5. The results are shown in Table 2. It turns out that the best improvement we were able to obtain is very small (0.5% of MAP increase). There may be several reasons for that. First, the accuracy of the automatic document tagging may be insufficient. Second, even if elements are correctly identified in documents, if queries are treated as bags-of-words then any PICO element can match with any identical word in the query, whether it describe the same element or not. However, we also tested a naïve approach that matches the PICO elements in queries with the corresponding elements in documents. But this approach quickly turns out to be too restrictive and leads to bad results.

| Measure | Weighted elements | | | |
|---|---|---|---|---|
| | P | I / C | O | Best† |
| MAP increase | 0.0% | −0.2% | −0.1% | +0.5% |

Table 2: Results using the PICO elements automatically detected in documents (†: $w_P = 0.5$, $w_I = 0.2$).

As we can see, this approach only brings limited improvement in retrieval effectiveness. This rises the question of the usability of such tagging method in its current performance state. We will see in the next section an alternative solution to this problem that relies on the distribution of PICO elements in documents.

## 4 Method

### 4.1 Distribution of PICO elements

PICO elements are not evenly distributed in medical documents, which often follow some implicit writing convention. An intuitive method is to weight higher a segment that is more probable to contain PICO elements. The distribution of PICO elements is likely to correlate to the position within the document. This intuition has been used in most of the supervised PICO detection methods which use location-based features. There has been several studies that cover the PICO extraction problem. However, as far as we know, none of them analyses and uses the positional ditribution of these elements within the documents for the purpose of IR. Biomedical abstracts can be typically represented by four ordered rhetorical categories which are Introduction, Methods, Results and Discussion (IMRAD) (Sollaci and Pereira, 2004). The reason is found in the need for speed when reviewing literature, as this format allows readers to pick those parts of particular interest. Besides, many scientific journals explicitly recommended this ordered structure.

The PICO dispersion is highly correlated to these rhetorical categories as some elements are more likely to occur in certain categories. For example, outcomes are more likely to appear in Results and/or Discussion parts. One could also expect to infer the

role played by PICO elements in a clinical study. For example, the drug *pioglitazone* has not the same role in a clinical study if it appears as the main intervention (likely to occur in all parts) or as a comparative treatment (Methods and/or Results parts).

Instead of analysing the dispersion of PICO elements into the four IMRAD categories, we choose to the use automatically splitted parts. There are several reasons for that. First, the IMRAD categories are not explicitely marked in abstracts. An automatic tagging of these would surely result in some errors. Second, using a low granularity approach would provide more precise statistics. Furthermore, if one would use the dispersion of elements as a criterion to estimate how important each part is, an automatic partition would be a good choice because of its repeatability and ease to implement.

We divided each manually annotated abstract into 10 parts of equal length (P1 being the begining and P10 the ending) and computed statistics on the number of elements than occur in each of these parts. The Figure 1 shows the proportion of elements for each part. We can observe that PICO elements are not evenly distributed throughout the abstracts. Universally accepted rules that govern medical writing styles would be the first reason for that. It is clear that the beginning and ending parts of abstracts do contain most of the PICO elements. This gives us a clear indication on which parts should be enhanced when searching for these elements.

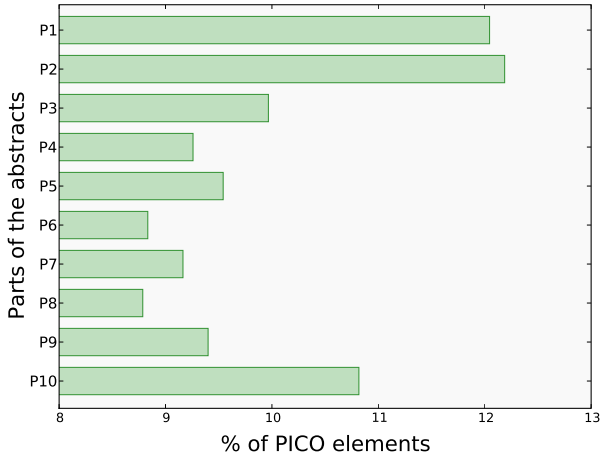


Figure 1: Proportion of PICO elements computed for each different part of abstracts.

Therefore, there may be several levels of granularity when using the PICO framework in IR. One can identify each PICO element in the document, whether it is described by a word, a phrase or a complete sentence. One can also use a coarser-grain approach, estimating from the distribution across documents the probability that each part contains a PICO element. As attempts to precisely locate PICO elements have shown that this task is particularly difficult, we propose to get rid this issue by using the second method.

### 4.2 Model definitions

We propose three different models that extend the classical language modeling approach. The first uses the structural information of documents, the second takes advantage of the PICO query structure while the third simply combine the first two models.

#### Model-1

Attempts to precisely locate PICO elements in documents have shown that this task is particularly difficult. We propose to get around this issue by introducing structural markers to convey document structure and use them as a means of providing location information. Accordingly, each document is represented as a series of successive parts. To integrate document structure into the ranking function, we estimate a series of probabilities that constraints the word counts to a specific part instead of the entire document. Each document $d$ is then ranked by a weighted linear interpolation. Intuitively, the weight of a part should depend on how much information is conveyed by its words. Given $\gamma_p$ the weight of the part $p \in [\text{TITLE}, P1 \cdots P10]$, $P(w \mid M_d)$ in equation 1 is re-defined as:

$$P_2(w \mid M_d) \propto P(w \mid M_d) + \sum_{p \in d} \gamma_p \cdot P(w \in p \mid M_d)$$

#### Model-2

The PICO formulation of queries provides information about the role of each query word. One idea is to use this structural decomposition to thoroughly balance elements in the ranking function. For example, the weight given to the drug *fluoxetine* should be different depending on whether it refers to the intervention or comparison concept. The same goes for *obesity* which can be a problem or an outcome. To

integrate this in the ranking function, we define a parameter $\delta_e$ that represents the weight given to query words belonging to the element $e \in$ [P, I, C, O]. $f(w, e) = 1$ if $w \in e$, 0 otherwise. We re-defined $P(w \mid M_d)$ in equation 1 as:

$$P_3(w \mid M_q) \propto P(w \mid M_q) + \sum_{e \in [P,I,C,O]} \delta_e \cdot f(w, e) \cdot P(w \mid M_q)$$

**Model-1+2**

This is the combination of the two previously described models. We re-defined the scoring function as:

$$score(q, d) = \sum_{w \in q} P_3(w \mid M_q) \cdot \log P_2(w \mid M_d)$$

## 5 Experiments

In this section, we describe the details of our experimental protocol. We then present the results obtained with the three proposed models.

**Experimental settings**

We gathered a collection of nearly 1.5 million abstracts from PubMed with the following requirements: with abstract, humans subjects, in english and selecting the following publication types: RCT, reviews, clinical trials, letters, practice guidelines, editorials and meta-analysis. Prior to the index construction, each abstract is automatically divided into 10 parts of equal length, abstracts containing less than 10 words are discarded. The following fields are then marked: TITLE, P1, P2, ... P10 with P1 being the begining of the document and P10 the ending.

Unfortunately, there is no standard test collection appropriate for testing the use of PICO in IR and we had to manually create one. For queries, we use the Cochrane systematic reviews[4] on 10 clinical questions about different aspects of "*diabetes*". These reviews contain the best available information about an healthcare intervention and are designed to facilitate the choices that doctors face in health care. All the documents in the "Included studies" section are judged to be relevant for the question. These included studies are selected by the reviewers (authors of the review article) and judged to be highly related to the clinical question. In our experiments, we consider these documents as relevant ones. From the 10 selected questions, professors in family medicine have formulated a set of 52 queries, each of which was manually annotated according to the PICO structure. The resulting testing corpus is composed of 52 queries (average length of 14.7 words) and 378 relevant documents. Below are some of the alternative formulations of queries for the question "*Pioglitazone for type 2 diabetes mellitus*":

*In patients with type 2 diabetes* (P) | *does pioglitazone* (I) | *compared to placebo* (C) | *reduce stroke and myocardial infarction* (O)

*In patients with type 2 diabetes who have a high risk of macrovascular events* (P) | *does pioglitazone* (I) | *compared to placebo* (C) | *reduce mortality* (O)

We use cross-validation to determine reasonable weights and avoid over-fitting. We have divided the queries into two groups of 26 queries: $Q_a$ and $Q_b$. The best parameters found for $Q_a$ are used to test on $Q_b$, and vice versa. In our experiments, we use the KL divergence ranking (equation 1) as baseline. The following evaluation measures are considered relevant:

*Precision at* $n$ ($P@n$). Precision computed on only the $n$ topmost retrieved documents.

*Mean Average Precision* (MAP). Average of precisions computed at the point of each relevant document in the ranked list of retrieved documents.

MAP is a popular measure that gives a global quality score of the entire ranked list of retrieved documents. In the case of clinical searches, one could also imagine this scenario: a search performed by a physician who does not have the time to look into large sets of results, but for whom it is important to have relevant results in the top 10. In such case, $P@10$ is also an appropriate measure.

Student's t-test is performed to determine statistical significance. The Lemur Toolkit[5] was used for

[4] www.cochrane.org/reviews/

[5] www.lemurproject.org

all retrieval tasks. Experiments were performed with an "out-of-the-box" version of Lemur, using its tokenization algorithm and porter stemmer. The Dirichlet prior smoothing parameter was set to its default value $\mu = 2500$.

### Experiments with model-1

We first investigated whether assigning a weight to each part of the document can improve the retrieval accuracy. It is however difficult to determine a set of reasonable values for all the parts together, as the value of one part will affect those of the others. In this study, we perform a two pass tuning. First, we consider the $\gamma_p$ weights to be independent. By doing so, searching for the optimal weight distribution can be seen as tuning the weight of each part separately. When searching the optimal weight of a part, the weight for other parts is assigned 0. Second, these approximations of the optimum values are used as initial weights prior to the second pass. The final weight distribution is obtained by searching for the best weight combination around the initial values.

The Figure 2 shows the optimal weight distributions along with the best relative MAP increase for each part. A noticeable improvement is obtained by increasing the weights associated to the title/introduction and conclusion of documents. This is consistent with the results observed on the distribution of PICO elements in abstracts. Boosting middle parts of documents seems to have no impact at all. We can see that the two $\gamma_p$ weight distributions (1-pass and 2-pass) are very close.

Performance measures obtained by model-1 are presented in Table 3. With 1-pass tuning, we observe a MAP score increase of 37.5% and a P@10 increase of 64.1%. After the second pass, scores are lower with 35% and 60.5% for MAP and P@10 respectively. This result indicates that there is possibly overfitting when we perform the two pass parameter tuning. It could also be caused by the limited number of query in our test collection. However, we can determine reasonable weights by tuning each part weight separately.

### Experiments with model-2

We have seen that a large improvement could come from weighting each part accordingly. In a second series of experiments, we try to assign a different

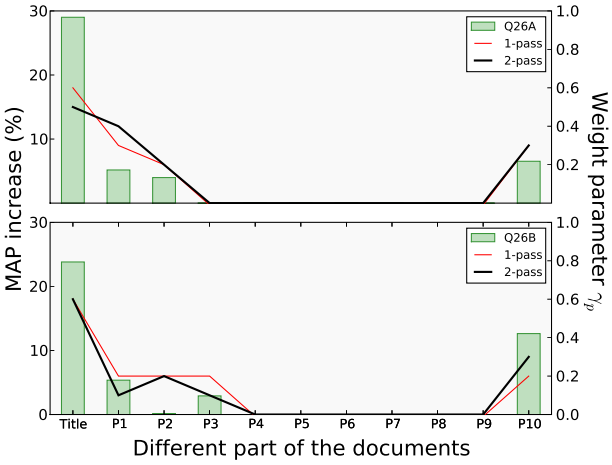


Figure 2: Best MAP increase for each part p (bar charts), corresponding 1 and 2-pass $\gamma_p$ weights are also given.

weight to each PICO element in queries. A grid search was used to find the optimal $\delta_e$ weights combination. The results are shown in Table 3.

We observe a MAP score increase of 22.5% and an increase of 11% in P@10. Though the use of a PICO weighting scheme increases the retrieval accuracy, there is clearly much to gain by using the document structure. The optimal $[\delta_p, \delta_i, \delta_c, \delta_o]$ weights distribution is $[0.3, 1.2, 0, 0.1]$ for Qa and $[0.2, 1, 0, 0.2]$ for Qb. That means that the most important words in queries belong to the Intervention element. This supports the manual search strategy proposed by (Weinfeld and Finkelstein, 2005), in which they suggested that I and P elements should be used first to construct queries, and only if too many results are obtained that other elements should be considered.

It is interesting to see that query words belonging to the Comparison element have to be considered as the least important part of a query. Even more so because they are in the same semantic group as the Intervention words. A reason for that could be the use of vague words such as "*no-intervention*" or "*placebo*". The methodology employed to construct the queries is also responsible. Indeed, physicians have focused on producing alternative formulations of 10 general clinical questions by predominantly modifying the one of the PICO elements. As a result, some of them do share the same vague Comparison words.

| Experiments | MAP | | | P@10 | | |
|---|---|---|---|---|---|---|
| | Qb→Qa | Qa→Qb | % Avg. | Qb→Qa | Qa→Qb | % Avg. |
| Baseline | 0.118 | 0.131 | | 0.219 | 0.239 | |
| Model-1 / 1pass | 0.165 | 0.176 | +37.5%‡ | 0.377 | 0.373 | +64.1%‡ |
| Model-1 / 2pass | 0.165 | 0.170 | +35.0%‡ | 0.354 | 0.381 | +60.5%‡ |
| Model-2 | 0.149 | 0.168 | +22.5%‡ | 0.250 | 0.258 | +11.0% |
| Model-1+2 | 0.198 | 0.202 | +61.5%‡ | 0.385 | 0.392 | +70.0%‡ |

Table 3: Cross-validation (train→test) scores for the baseline (Kullback-Leibler divergence), **model-1** with 1 and 2-pass tuning, **model-2** and their combination (**model-1+2**). Relative increase over the baseline is also given (averaged between Qa and Qb). (‡: t.test $< 0.01$)

**Experiments with model-1+2**

We have seen that both the use of a location-based weighting and a PICO-structure weighting scheme increase the retrieval accuracy. In this last series of experiments, we analyse the results of their combination. We can observe that fusing model-1 and model-2 allows us to obtain the best retrieval accuracy with a MAP score increase of 61.5% and a P@10 increase of 70.0%. It is a large improvement over the baseline as it means that instead of about two relevant documents in the top 10, our system can retrieve nearly four. These results confirm that both PICO framework and document structure can be very helpful for the IR process.

## 6 Conclusion

We presented a language modeling approach that integrates document and PICO structure for the purpose of clinical IR. A straightforward idea is to detect PICO elements in documents and use the elements in the retrieval process. However, this approach does not work well because of the difficulty to arrive at a consistent tagging of these elements. Instead, we propose a less demanding approach which assigns different weights to different parts of a document.

We first analysed the distribution of PICO elements in a manually annotated abstracts collection. The observed results led us to believe that a location-based weighting scheme can be used instead of a PICO detection approach. We then explored whether this strategy can be used as an indicator to refine document relevance. We also proposed a model to integrate the PICO information provided in queries and investigated how each element should be balanced in the ranking function. On a data set composed of 1.5 million abstracts extracted from PubMed, our method obtains an increase of 61.5% for MAP and 70% for P@10 over the classical language modeling approach.

This work can be much improved in the future. For example, the location-based weighting method can be improved in order to model a different weight distribution for each PICO element. As the distribution in abstracts is not the same among PICO elements, it is expected that differentiated weighting schemes could result in better retrieval effectiveness. In a similar perspective, we are continuing our efforts to construct a larger manually annotated collection of abstracts. It will be thereafter conceivable to use this data to infer the structural weighting schemes or to train a more precise PICO detection method. The focused evaluation described in this paper is a first step. Although the queries are limited to diabetes, this does not affect the general PICO structure in queries. We plan to extend the coverage of queries to other topics in the future.

## Acknowledgements

The work described in this paper was funded by the Social Sciences and Humanities Research Council (SSHRC). The authors would like to thank Dr. Ann McKibbon, Dr. Dina Demner-Fushman, Lorie Kloda, Laura Shea, Lucas Baire and Lixin Shi for their contribution in the project.

## References


A. Andrenucci. 2008. Automated Question-Answering Techniques and the Medical Domain. In *International Conference on Health Informatics*, volume 2, pages 207–212.

A.R. Aronson. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. In *AMIA Symposium*.

G. Chung. 2009. Sentence retrieval for abstracts of randomized controlled trials. *BMC Medical Informatics and Decision Making*, 9(1):10.

M. Dawes, P. Pluye, L. Shea, R. Grad, A. Greenberg, and J.Y. Nie. 2007. The identification of clinically important elements within medical journal abstracts: Patient-Population-Problem, Exposure-Intervention, Comparison, Outcome, Duration and Results (PECODR). *Informatics in Primary care*, 15(1):9–16.

D. Demner-Fushman and J. Lin. 2006. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *ACL*.

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

N. Fuhr and K. Großjohann. 2001. XIRQL: A query language for information retrieval in XML documents. In *SIGIR*, pages 172–180.

M.J. Hansen, N.O. Rasmussen, and G. Chung. 2008. A method of extracting the number of trial participants from abstracts describing randomized controlled trials. *Journal of Telemedicine and Telecare*, 14(7):354–358.

INEX. 2002-2009. Proceedings of the INitiative for the Evaluation of XML Retrieval (INEX) workshop.

J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. 2005. Structured queries in XML retrieval. In *CIKM*, pages 4–11.

J.M. Ponte and W.B. Croft. 1998. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281.

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

D.L. Sackett, W. Rosenberg, J.A. Gray, R.B. Haynes, and W.S. Richardson. 1996. Evidence based medicine: what it is and what it isn't. *British medical journal*, 312(7023):71.

C. Schardt, M. Adams, T. Owens, S. Keitz, and P. Fontelo. 2007. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics and Decision Making*, 7(1):16.

L.B. Sollaci and M.G. Pereira. 2004. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the Medical Library Association*, 92(3):364.

A. Trotman. 2005. Choosing document structure weights. *Information Processing and Management*, 41(2):243–264.

J.M. Weinfeld and K. Finkelstein. 2005. How to answer your clinical questions more efficiently. *Family practice management*, 12(7):37.

R. Wilkinson. 1994. Effective retrieval of structured documents. In *SIGIR*, pages 311–317.