# An Ensemble Model of Clinical Information Extraction and Information Retrieval for Clinical Decision Support

Yanshan Wang, Majid Rastegar-Mojarad, Ravikumar Komandur-Elayavilli,
Sijia Liu and Hongfang Liu

Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA
{Wang.Yanshan, Mojarad.Majid, KomandurElayavilli.Ravikumar, Liu.Sijia,
Liu.Hongfang}@mayo.edu

**Abstract.** This paper describes the participation of Mayo Clinic NLP team in the Text REtreival Conference (TREC) 2016 Clinical Decision Support track. We propose an ensemble model which combines three components: a Part-of-Speech based query term weighting model (POS-BoW); a Markov Random Field model leveraging clinical information extraction (IE-MRF); and a Relevance Pseudo Feedback (RPF) model. We submitted three automatic runs and two manual runs. The experimental results show that the automatic runs outperform the median results of all participant teams for up to 76.7% of the given query topics.

**Keywords:** ensemble model, information extraction, information retrieval, Part-of-Speech, Markov Random Field

## 1 Introduction

Text REtreival Conference 2016 Clinical Decision Support (TREC 2016 CDS) track focuses on biomedical literature retrieval that helps physicians find the precise literature information and make the best clinical decision at the point of care. 1.25 million articles from PubMed Central (PMC) are used as the document collection, which is a subset of open access articles from PMC on March 28, 2016. Electronic Health Records (EHRs) from MIMIC-III data set [4] were utilized to generate the query topics. Those topics are categorized into three most common types, namely *Diagnosis*, *Test* and *Treatment*, according to physicians' information needs, and 10 topics are provided for each type. Each topic is comprised of a *note* field (admission note), a *description* field (jargons and clinical abbreviations are removed) and a *summary* field (simplified version of the description). The participants are required to use only one of these three fields in their submissions and at least one submission must utilize the *note* field. Submitted systems should retrieve relevant biomedical articles for each given query topic to answer three corresponding clinical questions: *What is the patient's diagnosis? What tests should the patient receive? How should the patient be treated?*.

Each participant was allowed to submit up to 5 runs with up to 1000 documents per query.

We propose an ensemble model which combines three models: a Part-of-Speech based query term weighting model (POS-BoW), a Markov Random Field model leveraging clinical information extraction (IE-MRF), and a Relevance Pseudo Feedback (RPF) model for query expansion. The POS-BoW model is a revised bag-of-words (BoW) model, which assigns weights to query terms according to Part-of-Speech (POS) [11,12]. IE-MRF applies Markov Random Field (MRF) model [6] to the medical concepts extracted by an automatic clinical information extraction method. RPF utilized co-occurred MeSH headings to expand the query topics. Each of the constituent model embeds one kind of information from the topics into the final topic representations: POS-BoW considers the original query topics; IE-MRF leverages the medical concepts in the topics; and RPF utilizes the inner connection between topics and articles. We submitted five runs to the TREC 2016 CDS track including three automatic runs and two manual runs. Both automatic runs and manual runs used the ensemble model but differently the manual runs utilized the medical concepts in the MRF model that were manually extracted by an expert with medical background.

The rest of this paper is organized as follows. Section 2 describes the methodology details in our submissions. The experiments and results are shown in Section 3. Section 4 concludes our study.

## 2 Methods

Among the five runs that we submitted to the TREC 2016 CDS track, two runs used the topic *note* field (one manual run, denoted as 'mayomn' and one automatic run, denoted as 'mayoan'), two runs used the topic *description* field (one manual run, denoted as 'mayomd' and one automatic run, denoted as 'mayoad') and one run used the topic *summary* field (one automatic run, denoted as 'mayoas'). The submitted runs are summarized in Table 1. In this section, we describe the methodologies used in the submitted runs.

Table 1: Summary of submitted runs

| Run Name | Method | Query Section |
|---|---|---|
| mayomn | manual | note |
| mayoan | automatic | note |
| mayomd | manual | description |
| mayoad | automatic | description |
| mayoas | automatic | summary |

### 2.1 Ensemble Model

Fig. 1 shows the structure of ensemble model, i.e., how the constituent models are combined into the ensemble model. Briefly, the POS-BoW model generates weighted queries, the RPF model generates expanded queries, and the MRF model generates weighted medical concepts by utilizing the extracted medical concepts. Finally we linearly combine these constituent models by assigning different weights which are carefully tuned on the TREC 15 CDS track. In the following subsections, we will describe the details of these constituent models.

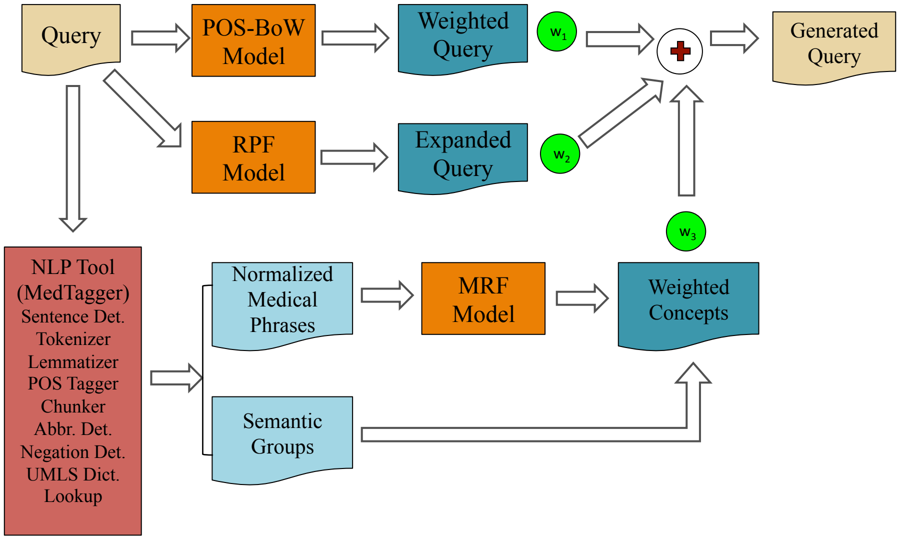


Fig. 1: Pipeline of query generation.

### 2.2 Part-of-Speech based Query Term Weighting (POS-BoW)

The BoW model assumes that a document consists of an unordered set of independent words. Many sophisticated information retrieval models were developed based on the BoW assumption [8,7,3,13,10]. In our study, we utilized the POS-BoW model which assigns different weights to query terms according to the POS [12]. POS is an important indicator for the term informativeness, particularly in medical domain. For example, Topic 19 of the TREC 2016 CDS track “atenolol was switched to metoprolol”, “atenolol” and “metoprolol” are proper nouns which are the most informative terms to understand the sentence semantics while “switched” is a past participle verb which describes the action between two proper nouns. Thus, different weights should be assigned to the proper nouns and the past participle verbs.

How to determine the weight for each POS category is a big challenge. Wang et.al [12] proposes a machine learning algorithm to train the weights for seven POS categories (*singular or mass nouns (NN), plural nouns (NNS), past participle verbs (VBN), past tense verbs (VBD), adjectives (JJ), adverbs (RB), singular proper nouns (NNP)*) based on the TREC 2011 and 2012 Medical Records tracks. Since the topics given in the TREC 2016 CDS are also generated from EHRs, we utilize the trained weights from [12] in the submitted runs. The weights are::

$$\begin{aligned}&Weight\{NN, NNS, VBN, VBD, JJ, RB, NNP\} = \\ &\qquad \{0.5970, 0.2265, 0.3065, 0.2260, 0.3730, 0.1040, 0.8930\}.\end{aligned}$$

As an example, the summary field of Topic 1 in the TREC 2016 CDS track: "A 78 year old male presents with frequent stools and melena." can be represented as follows in Indri [9]:

```
#weight( 0.0 a 0.0 78 0.597 year 0.373 old 0.597 male 0.0 presents
 0.0 with 0.373 frequent 0.2265 stools 0.0 and 0.597 melena)
```

### 2.3 Clinical Information Extraction Enhanced Markov Random Field Model (IE-MRF)

Apart from the unigram query terms retained in the previous POS-BoW model, the medical concepts are essential for understanding medical semantics. MedTagger [5] is utilized to extract and normalize the medical concepts based on Unified Medical Language System (UMLS). For each medical concept, the MRF model is used to model term dependencies [6]. Three variants of the MRF model, i.e., full independence, sequential dependence and full dependence, are implemented with weights 0.4, 0.35 and 0.4, respectively. A weight of 2.0 is given to each MRF generated multi-word medical concept while 1.0 is given to each single-word medical concept.

We employed MedTagger to extract the semantic groups for each medical concept. For different topic types, i.e., *Diagnosis*, *Test* and *Treatment* topics, the semantic group might indicate the significance of concepts. Therefore, different weights were considered for the medical concepts based on the semantic group and topic type. Table 2 lists the weights used in our systems. We note that all the weights are obtained by tuning on the TREC 15 CDS track. For example, a medical abbreviation "CBD" in the summary field of Topic 6 is firstly extracted and subsequently normalized as "common bile duct" by MedTagger; then, the MRF model formulates this medical concept as follows in Indri:

```
#weight( 2.0 #weight(
              0.4  #combine( common bile duct )
              0.35 #combine( #1(bile duct)  #1(common bile)
                             #1(common bile duct) )
              0.4  #combine( #uw8(bile duct)  #uw8(common duct)
                             #uw8(common bile)  #uw8(common bile duct) ) )
        )
```

Table 2: Weights for concepts based on different semantic groups and different query types

| Semantic Groups | *Diagnosis* | *Test* | *Treatment* |
|---|---|---|---|
| ANAT | 2 | 1 | 1 |
| CHEM;DRUG | 1 | 1 | 2 |
| DISO | 2 | 1 | 1 |
| DRUG | 1 | 1 | 2 |
| FIND | 2 | 1 | 1 |
| PROC | 2 | 1 | 3 |
| others | 1 | 2 | 1 |

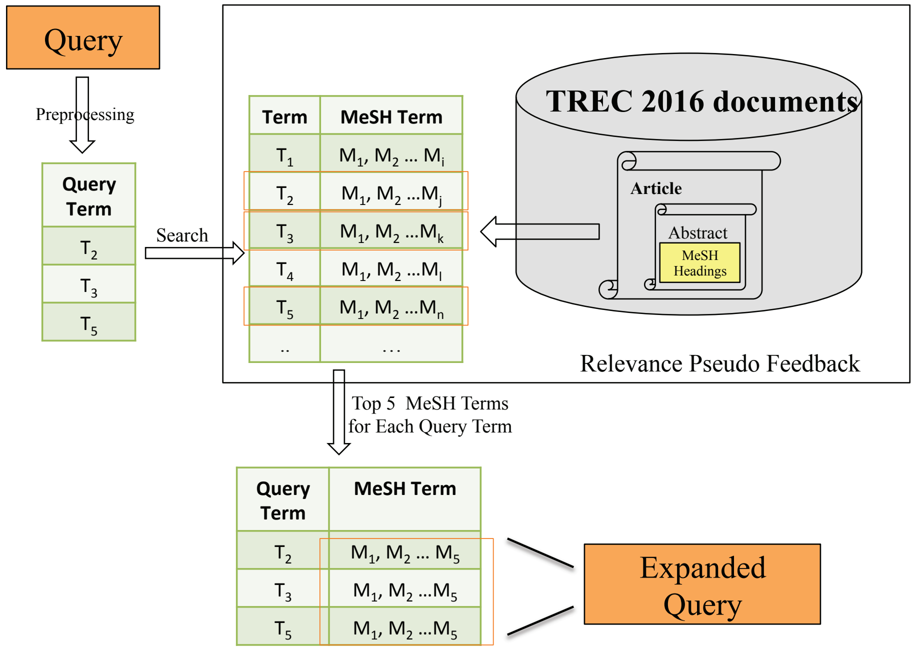

Query
Preprocessing
Query Term
$T_2$
$T_3$
$T_5$
Search
Term
MeSH Term
$T_1$ $M_1, M_2 \dots M_i$
$T_2$ $M_1, M_2 \dots M_j$
$T_3$ $M_1, M_2 \dots M_k$
$T_4$ $M_1, M_2 \dots M_l$
$T_5$ $M_1, M_2 \dots M_n$
.. …
TREC 2016 documents
Article
Abstract
MeSH Headings
Relevance Pseudo Feedback
Top 5 MeSH Terms for Each Query Term
Query Term
MeSH Term
$T_2$ $M_1, M_2 \dots M_5$
$T_3$ $M_1, M_2 \dots M_5$
$T_5$ $M_1, M_2 \dots M_5$
Expanded Query

Fig. 2: Pipeline of Relevance Pseudo Feedback.

### 2.4 Query Expansion using Relevance Pseudo Feedback (RPF)

To expand queries, we took advantage of inner connection between query terms and PubMed articles. Specifically, we used the idea behind pseudo relevance feedback [2] and developed a Relevance Pseudo Feedback (RPF) model. Figure 2 illustrates the pipeline of RPF model. Instead of extracting expansion terms from top ranking retrieved documents, we utilized MeSH Headings, manually assigned by experts to each MEDLINE abstracts, to expand queries. We considered MeSH Headings as relevance feedback and created a list of correlated (term, MeSH Heading) pairs based on their co-occurrence in MEDLINE abstracts. In order to identify these pairs, first using Eutils API [1], we retrieved MEDLINE articles metadata, containing title, abstract, MeSH headings, etc. Then, we generated a list of terms appeared in the titles and abstracts. After removing stop-words from the list, we counted the number of co-occurrence of each term and MeSH headings and used it to calculate odd ratio for each (term, MeSH heading) pair. To expand each query, first we split the query into terms and then for each term, 5 top correlated MeSH headings (raked by odd ratio), are added to the original query. For example, the expansion for the summary field of Query 1 in the TREC 2016 CDS track are:

```
#combine( copulation plant infertility circumcision male loss of
heterozygosity chromosomes human pair 3 tension type headache
poliomyelitis feces diarrhea humans male )
```

### 2.5 Manual Runs

In the manual runs, we also applied the ensemble model but differently we utilized the medical concepts extracted by a domain expert instead of MedTagger. The manual generation of medical concepts essentially consists of two steps:

1. The query topics were pre-processed to identify the shallow chunks of the text, such as base noun phrases (without any preposition attachments) and verb phrases.
2. The domain expert carefully analyzed these linguistic phrases chunks and selected only those phrases that are relevant after ignoring the irrelevant phrases. The expert also removed certain words from the linguistic phrases based on the judgement whether it would be a noise and identify false positives.

In addition, the expert supplemented the phrases with additional concepts in order to improve the recall of the system. Query expansion was done based on two criteria:

1. Translating inferences derived from quantitative lab values to qualitative concepts. Consider the following example: “Her hematocrit dropped from 28 to 16.”. The domain expert translated this concept to ”hematocrit drop” in the final query. Similarly the domain expert while considering the concept “elevated LDH to 315” augmented the search query with the concept“Hyperlipidemia” based on the lab value of LDL in addition to the “elevated LDL”.
2. Inferences drawn based on context analysis derived from PubMed Search results. Consider the example “Guaiac was reported as being positive”. Searching “Guaiac” in PubMed yielded other related concepts, such as “colorectal cancer”, “fecal immunochemical test” etc., which were added to the queries.

[1] https://www.ncbi.nlm.nih.gov/books/NBK25501/

For the first criterion only the content in the given query was considered, while in the second additional contextual information such as concepts that are semantically related to phrases generated from the query were considered. While the aim of former was to retrieve precise results the goal of latter was to improve the recall without sacrificing precision.

## 3 Experiments and Results

Indri [9] was utilized as our indexing and retrieval tool. The preprocessing included stopword removal and Porter stemming. The stopword list was based on the PubMed stopwords [2]. The *article-id*, *title*, *abstract* and *body* fields of each document were indexed. Language models with two-stage smoothing [14] was used to obtain all the retrieval results.

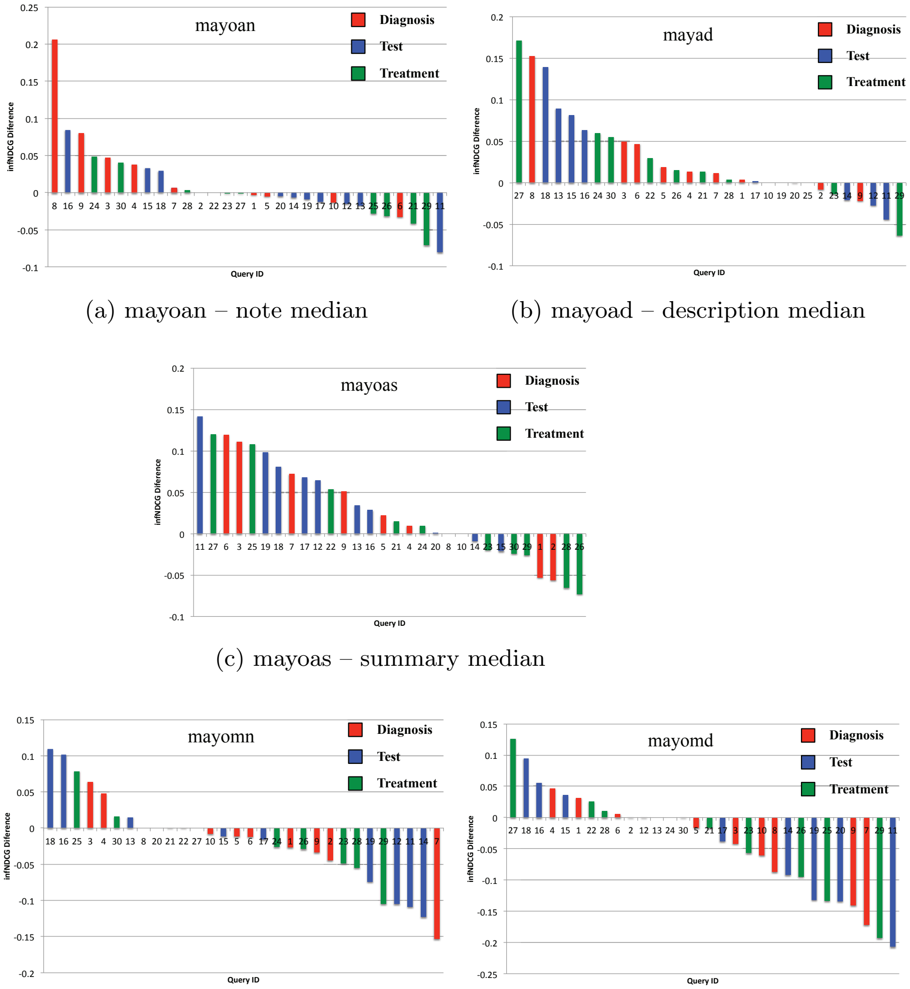
Table 3 summarizes the results of submitted runs in terms of Inferred Normalized Discounted Cumulated Gain (infNDCG), R-precision (R-prec) and Precision at 10 (P@10). As shown in Table 3, 'mayoas', an automatic run using topic summaries, performs the best among all submitted runs. 'The reason might be that for machine systems, summary texts that capture the main topics in queries are less noisy than descriptive texts. The performance of manual run based on queries from topic notes ('mayomn') is better than the queries from topic descriptions ('mayomd'). This may be due to the fact that the notes being more concise than descriptive texts are more informative and less noisy, from which humans are able to infer more knowledge. Interestingly, we observe that 'mayoad', an automatic run using topic description, outperforms 'mayomd'.

By comparing the medical concepts extracted by the NLP software and the human expert, we found that the inference drawn by the human expert might be the key factor that resulted in inferior performance. For example, "coronary artery bypass" was inferred by the human expert for Query 1, which was never mentioned in the query context. We need a further study to evaluate the impact of inference, either by human expert or automatic software, for clinical decision making and clinical information retrieval. As being studied, clinical decision can be dependent upon the physician's ability to reason, think, and judge [1].

Table 3: Results of all submitted runs

| Run Name | infNDCG | R-prec | P@10 |
|---|---|---|---|
| mayomn | 0.1407 | 0.1042 | 0.2167 |
| mayoan | 0.1309 | 0.0985 | 0.2200 |
| mayomd | 0.1199 | 0.0832 | 0.1600 |
| mayoad | 0.1315 | 0.0975 | 0.2167 |
| mayoas | **0.2146** | **0.1659** | **0.3067** |

[2] http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/



(a) mayoan – note median

(b) mayoad – description median

(c) mayoas – summary median

(d) mayomn – manual median

(e) mayomd – manual median

Fig. 3: Difference of infNDCG between the submitted runs and the median results of all participant teams.

For each query topic, we calculate the difference of infNDCG between our submitted runs and the median results of corresponding runs by all participant teams and illustrate the results in Figure 3. Out of 30 query topics, the submitted automatic runs, 'mayoan', 'mayoad' and 'mayoas', have better performance than the median for 13 (43.3%), 23 (76.7%), and 21 (70.0%) query topics, respectively. Unlike automatic runs, the submitted manual runs, 'mayomn' and 'mayomd', only perform better for 12 (40.0%) and 14 (46.7%) out of 30 topics than manual median results.

In our experiments, we implemented an information extraction component to extract age and gender related information from titles, abstracts, and queries and utilized the information to re-rank retrieved documents. In addition, we also tried extracting and searching UMLS concepts using Concept Unique Identifiers (CUIs) for both query topics and corpus. However, neither of the two approaches could improve the results when tested on TREC 15 CDS track. Thus, they were not utilized in the submitted runs.

## 4 Conclusion

In this paper we present the system we developed in the participation of TREC 2016 CDS track. We submitted five runs that consist of three automatic runs and two manual runs using different fields in the provided query topics. In the automatic runs, we utilized an ensemble model that combines three sophisticated methods. The automatic runs 'mayoad' and 'mayoas' outperform the median result in 76.7% and 70.0% of topics, respectively. Among the 26 participant teams, our best automatic runs ranked 10th and 5th in terms of infNDCG and P@10, respectively.

## Acknowledgments

The authors gratefully acknowledge the support from the National Library of Medicine (NLM) grant R01LM11934.

## References

1. Benner, P., Hughes, R.G., Sutphen, M.: Clinical reasoning, decisionmaking, and action: Thinking critically and clinically (2008)
2. Efthimiadis, E.N.: Query expansion. Annual review of information science and technology 31, 121–187 (1996)
3. Jelinek, F.: Interpolated estimation of markov source parameters from sparse data. Pattern recognition in practice pp. 381–402 (1980)
4. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data 3 (2016)

5. Liu, H., Bielinski, S.J., Sohn, S., Murphy, S., Wagholikar, K.B., Jonnalagadda, S.R., Ravikumar, K., Wu, S.T., Kullo, I.J., Chute, C.G.: An information extraction framework for cohort identification using electronic health records. AMIA Summits on Translational Science Proceedings 2013, 149 (2013)
6. Metzler, D., Croft, W.B.: A Markov random field model for term dependencies. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 472–479. ACM (2005)
7. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M., et al.: Okapi at trec-3. NIST SPECIAL PUBLICATION SP pp. 109–109 (1995)
8. Salton, G., McGill, M.J.: Introduction to modern information retrieval. McGraw-Hill, Inc. (1986)
9. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language model-based search engine for complex queries. In: Proceedings of the International Conference on Intelligent Analysis. vol. 2, pp. 2–6. Citeseer (2005)
10. Wang, Y., Lee, J.S., Choi, I.C.: Indexing by latent dirichlet allocation and an ensemble model. Journal of the Association for Information Science and Technology 67(7), 1736–1750 (2016), http://dx.doi.org/10.1002/asi.23444
11. Wang, Y., Wu, S., Li, D., Liu, H.: Influence of part-of-speech on the clinical information retrieval. In: AMIA iHealth Clinical Informatics Conference (2016)
12. Wang, Y., Wu, S., Li, D., Mehrabi, S., Liu, H.: A part-of-speech term weighting scheme for biomedical information retrieval. Journal of Biomedical Informatics 63, 379 – 389 (2016), http://www.sciencedirect.com/science/article/pii/S1532046416301125
13. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 334–342. ACM (2001)
14. Zhai, C., Lafferty, J.: Two-stage language models for information retrieval. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 49–56. ACM (2002)