Using Discharge Summaries to Improve Information Retrieval in Clinical Domain

Dongqing Zhu[1], Wu Stephen[2], Masanz James[2],
Ben Carterette[1], and Hongfang Liu[2]

[1] University of Delaware, 101 Smith Hall, Newark, DE 19716, USA
{zhu,carteret}@cis.udel.edu
[2] Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA
{wu.stephen,masanz.james,liu.hongfang}@mayo.edu

Abstract. Task 3 of the 2013 ShARe/CLEF eHealth Evaluation Lab simulated web searches for health information by patients. The web searches were designed to be connected to hospital discharge summaries from the patient's Electronic Medical Record (EMR), thus effectually modeling a post-visit information need. We primarily investigated three research questions about the retrieval of medical information from the web: 1) to what degree retrieval techniques effective in searching Electronic Medical Records (EMRs) could aid in finding medical web documents; 2) to what degree medical web retrieval would benefit from natural language processing (NLP) techniques that extract information from text based on medical knowledge; and 3) how to leverage contextual information in the patient's discharge summaries to improve retrieval. We submitted seven runs to ShARe/CLEF eHealth. Our best run used effective EMR-based IR techniques, NLP-produced information, and information in patients' discharge summaries to achieve precision at 10 (P@10) scores at or above the CLEF median for all but 2 of 50 test queries.

Keywords: language models, mixture of relevance models, Markov Random Field, semantic concepts, UMLS Metathesaurus, MeSH

1 Introduction

Task 3 of the 2013 ShARe/CLEF eHealth Evaluation Lab [1] simulated web searches for health information by patients. The web searches were designed to be connected to hospital discharge summaries from the patient's Electronic Medical Record (EMR), thus effectually modeling a post-visit consumer information need. This constrains the ad-hoc retrieval of webpages to a limited number of medical pages; it also differs from recent medical retrieval tasks like patient cohort identification from EMR text [2] in that it is web search.

This paper describes our participation in Task 3 of 2013 ShARe/CLEF eHealth Evaluation Lab. For this retrieval task, we primarily investigated three research questions: 1) to what degree retrieval techniques effective in searching Electronic Medical Records (EMRs) could aid in finding medical web documents; 2) to what degree medical web retrieval would benefit from natural language processing (NLP) techniques that extract information from text based on medical knowledge; and 3) how to leverage contextual information in the patient's discharge summaries to improve retrieval.

In particular, we used a two-step ranking strategy. In Step 1, we performed retrieval in the text space, where the documents and queries were in their raw text form, by using the MRF, MRM, and MeSH-based query expansion — the first research question. In Step 2, we re-ranked the output from Step 1 in the concept space where everything was represented by the medical concepts. In addition, we produced contextual artifacts of the medical concepts in documents and discharge summaries by using a clinical NLP annotation tool called MedTagger, and incorporated those artifacts in our retrieval models.

To derive the MRM model, we used several external sources: the TREC 2011 Medical Records Track test collection, the TREC 2007 Genomics Track test collection, and a subset of Mayo Clinic clinical notes collected between 2001–2010. To apply domain knowledge and construct the concept space, we adopted the Concept Unique Identifier (CUI) in the Unified Medical Language System (UMLS) for representing medical concepts.

We submitted seven runs to ShARe/CLEF eHealth. When evaluated by precision at 10 (P@10), our best run, which used the two-step ranking strategy and leveraged information in patients' discharge summaries, were above or at the CLEF median for all but 2 of 50 test queries.

The rest of the paper proceeds as follows: Sect. 2 presents the overview of our system and follows it with detailed description about the underlying retrieval models and medical-specific features. Then, Sect. 3 describes different system settings for our submitted runs, and shows the official evaluation results. Finally, Sect. 4 concludes our research findings.

2 Retrieval System

In this section, we first present an overview of our retrieval system, and then detail its underlying retrieval models and domain-specific features.

2.1 System Overview

Fig. 1 gives an overview of our complete system in its most complex form (run 4, see Sect. 3). The whole pipeline consists of six steps as indicated by circled numbers in the figure, and we describe each step below:

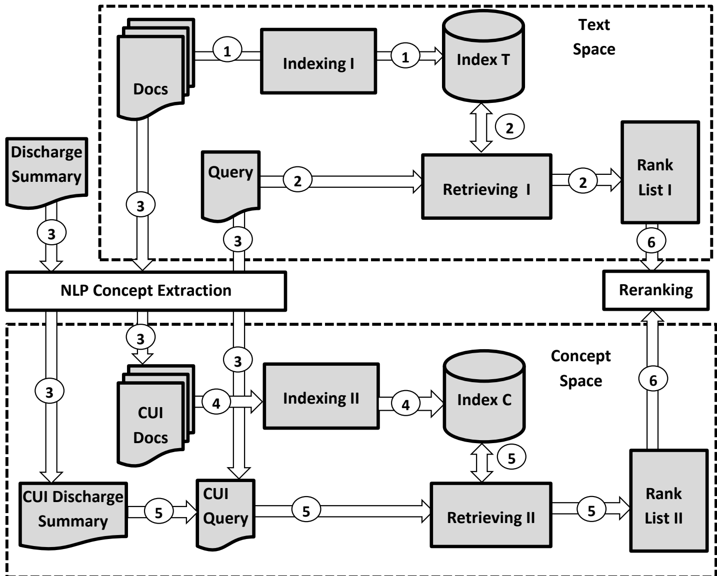


Fig. 1. System Overview

1. Indexing I: we used Indri[3] [3] to build the index T for the target collection (i.e., the collection of CLEF shared task) in the text space. In particular, we formulated each query based only on the information of the 'title' field in the query file. After html cleaning with Boilerpipe[4] [4], we stemmed words in both text documents and queries by Porter stemmer, and used a standard medical stoplist [5] for stopping words in queries only.

2. Retrieving I: we obtained intermediate retrieval results (i.e., rank list 1 in Fig. 1) by querying against the index T. We will detail the retrieval models in Sect. 2.2.

3. NLP Concept Extraction: we used MedTagger [6] to convert text into the concept space, i.e., CUIs. MedTagger extracts and annotates concepts from clinical text by leveraging large corpus statistics, machine learning, knowledge bases, and syntactic parsing. The output of MedTagger includes CUIs as well as their attributes, such negation, uncertainty, semantic group, etc. Documents in rank list 1, text queries, and the discharge summaries associated with each query were all mapped into the concept space in this manner.

4. Indexing II: we built another index for the CUI documents, again using Indri. However, we did not stem the CUIs since they are all unique. Due to the computational cost of NLP concept extraction on 1.6 million documents, we

[3] `http://www.lemurproject.org/indri/`
[4] `http://code.google.com/p/boilerpipe/`

only indexed the 61,273 documents that were present in retrieved results of the 50 test queries.

5. Retrieving II: the query likelihood model was used to retrieve CUI-only documents based on CUI-only queries. The concepts from discharge summaries were also included for the relevant submissions. We will elaborate on this in Sect. 2.3 where we describe our retrieval models in the concept space.
6. Re-ranking: we weight the text-based retrieval rankings against the concept-based retrieval rankings. Because retrieving in the concept space is done on the intermediate text-based output (Rank List 1), it is essentially a re-ranking procedure.

2.2 Text-based Retrieval

In this section, we describe our text-based retrieval models.

Query Likelihood Model. We used the query likelihood language model [7] as our baseline. This model scores documents for queries with the intuition that a user will think of the words in a document and try to write a query that is similar. Formally, the scoring function is a sum of the logarithms of smoothed probabilities:

$$\mathrm{score}(D, Q) = \log P(Q|D) = \sum_{i=1}^{n} \log \frac{\mathrm{tf}_{q_i,D} + \mu \frac{\mathrm{tf}_{q_i,C}}{|C|}}{|D| + \mu}, \tag{1}$$

where q_i is the ith term in query Q, n is the total number of terms in Q, $|D|$ and $|C|$ are the document and collection lengths in words respectively, $\mathrm{tf}_{q_i,D}$ and $\mathrm{tf}_{q_i,C}$ are the document and collection term frequencies of q_i respectively, and μ is the Dirichlet smoothing parameter. Smoothing is a common technique for estimating the probability of unseen words in the documents [8, 9]. An Indri query using this query likelihood model looks like the following:

```
#combine(shortness breath swelling).
```

Note that we formulated the above query by removing the stop words ‘of’ and ‘and’ from its original version ‘shortness of breath and swelling’.

This model has been shown to be effective in web search and in medical records; here, we test the intersection of the two domains — medical web texts. State-of-the-art performance on medical records is possible with language models when augmented with advanced retrieval models, which we turn to now.

Markov Random Field Model. In the query likelihood model, it is a strong assumption that query terms are generated independently from the document language model. In reality, related terms are likely to occur in close proximity to each other. The Markov random field (MRF) model [10] improves upon query likelihood model by incorporating term proximity information. It works by first constructing a graph that contains a document node, one node per query term, and edges that represent dependencies among nodes. Then, MRF models the joint distribution over the document random variable and query term random variables. The ranking function of the MRF model is of the form:

$$P_{\Lambda}(Q|D) \stackrel{\text{rank}}{=} \sum_{c \in T} \lambda_T f_T(c) + \sum_{c \in O} \lambda_O f_O(c) + \sum_{c \in O \cup U} \lambda_U f_U(c), \tag{2}$$

where T is defined to be the set of 2-cliques containing the document node and a query term node, O is the set of cliques involving the document node and two or more query terms that appear contiguously in the query, and U is the set of cliques involving the document node and two or more query terms that appear non-contiguously within the query. $f(c)$ is the feature function over clique c and λ's are the feature weights. MRF model has been shown to consistently outperform the standard unigram model across a range of TREC test collections [10, 11].

Following Metzler and Croft [10], we set the feature weights $(\lambda_T, \lambda_O, \lambda_U)$ to (0.8, 0.1, 0.1). An Indri query using MRF model looks like:

```
#weight( 0.8 #combine(shortness breath swelling)
 0.1 #combine( #1(breath swelling) #1(shortness breath) )
 0.1 #combine( #uw8(breath swelling)  #uw8(shortness breath) ) ),
```

where each '#1()' phrase specifies the ordered query terms, and each '#uw8()' phrase include two query terms that can occur within a text span of 8 terms.

Mixture of Relevance Models. We also used the mixture of relevance models (MRM) [12] to expand the query with related terms for reducing the vocabulary gap between query language and document language. Previous work has shown that MRM model significantly improve the retrieval in general web domain [12] as well as in the clinical domain [13, 14].

In particular, we selected several external collections along with the target collection (i.e., CLEF) for building our MRM model. Our preliminary work suggests that the domain and characteristics of the supporting external collections have an effect on retrieval performance, thus we choose the medically oriented resources in Table 1, which are shown with their collection statistics. The Medical and Genomics collections are from TREC 2011 Medical Records Track [2] and TREC 2007 Genomics Track [15] respectively. The MayoClinic collection is a subset of Mayo Clinic clinical notes collected between 2001–2010, retrieved from the Mayo Clinic Life Sciences System (MCLSS). This includes data from a comprehensive snapshot of Mayo Clinic's service areas, excluding only microbiology, radiology, ophthamology, and surgical reports. This corpus has been characterized for its clinical information content (namely, medical concepts [16] and terms [17]) and compared to other corpora, such as the 2011 MEDLINE/PubMed Baseline Repository and the 2010 i2b2 NLP challenge dataset [18].

Table 1. Collection Statistics

Collection	# of documents	Vocabulary size	Avg doc length
CLEF	1,628,823	10^6	892
Medical	100,866	10^5	423
Genomics	162,259	10^7	6,595
MayoClinic	39,449,222	10^6	346

Following the previous work [19], We weighted each expansion collection equally. An Indri query using MRM model looks like:

```
#weight( 0.6 #combine(shortness breath swelling)
 0.1 #weight(0.2 breath 0.19 hydroxocobalamin 0.12 short
  0.1 croup 0.09 asthma 0.07 chest 0.06 hydrocortison)

 0.1 #weight(0.22 breath 0.1 sleep 0.1 swell 0.09 airwai 0.09
  respiratori 0.09 clc 0.07 intub)

 0.1 #weight(0.19 swell 0.12 sudden 0.11 breath 0.096 feet
  0.09 frothi 0.07 short 0.05 swollen)

 0.1 #weight(0.11 failur 0.10 congest 0.09 swell 0.07
  chronic 0.07 fibril 0.07 atrial 0.07 pulmonari) ),
```

where the '#combine()' phrase corresponds to the original query, and the four inner '#weight()' phrases correspond to the relevance models derived respectively from four individual expansion collections listed in Table 1.

Query Expansion with MeSH. Previous work has shown that expanding queries with related Medical Subject Headings (MeSH) enhances retrieval performance significantly in an EMR cohort identification setting [19, 13]. Furthermore, since MeSH terms were designed to categorize scientific and medical literature, we believe MeSH is a good fit for the problem of searching the web for medical reference information. Thus, for each MetaMap-detected MeSH concept [20] in the query we included its entry terms and descendant nodes in the MeSH hierarchy for expansion [19]. An Indri query using MeSH-based query expansion looks like:

```
#weight( 0.7 #combine(shortness breath swelling)
```

```
0.3 #weight( 0.1 #uw16(dyspnea paroxysmal) 0.1 edema
0.1 hydrops 0.1 #uw16(edema cardiac) 0.1 #1(hydrops fetalis)
0.1 anasarca 0.1 dropsy 0.1 #1(shortness breath)
0.1 dyspneas 0.1 #1(breath shortnesses)) ),
```

where the '#combine()' phrase contains the original query, and the inner '#weight()' phrase contains MeSH expansion terms associated with MeSH terms detected in the original query.

2.3 Concept-based Retrieval for Re-ranking

In this section, we describe several concept-based retrieval models. We used these models for re-ranking the output from the text-based retrieval. The motivation of this re-ranking process was our second research question, namely, whether NLP approaches for concept identification could incorporate some medical domain knowledge and further improve the results from text-based retrieval.

Concept Unique Identifiers. In the concept space as shown in Fig. 1, we represented everything by UMLS CUI. These shallow semantics are a common normalization target for medical NLP systems, since extensive work has gone into the curation of the UMLS Metathesaurus in reconciling between many different source ontologies, terminologies, and vocabularies. We used the query likelihood model (see Eq. 1) for computing the relevance score. An Indri query using this model will look like:

```
#combine(C0225386 C0347940),
```

where the C0225386 is the CUI related to breath or respiratory air, and C0347940 is related to chest swelling.

Discharge Summary. Since queries were generated based on discharge summaries (DS), we hypothesized that DS could contain 'hidden' concepts that did not appear in the query but were related to query concepts. Psychologically, if a patient finishes reading his/her medical record, these 'hidden' concepts might serve as semantically priming for the patient's mind when he/she formulates the query. This may be valuable context that informs us about the patient's actual information need.

To identify these 'hidden' concepts and use them for query expansion, we again used MedTagger to convert text DS to CUI DS, and then weighted each CUI by its term frequency, and finally took the top 20 CUIs with their weights for query expansion. An Indri query formulated based on CUI DS looks like:

```
#weight( 0.8 #combine(C0225386 C0347940)
 0.2 #weight(37 C0018787 22 C0003842 21 C0397581 21 C0226004
```

```
20 C1269008 20 C0869781 20 C0205042 12 C0018802 11 C1278960
11 C0729538 11 C0398102 11 C0392907 11 C0042449 11 C0036186
10 C0030685 9 C2926602 9 C0460139 9 C0234222 9 C0012621
8 C1548828 ) ),
```

where expansion CUI terms (e.g., those related to heart and artery: C0018787, C0003842, and C0205042) are potentially useful for retrieving information related to chest swelling and shortness of breath.

Attributes. In addition to content artifacts (CUIs), Medtagger also produced contextual attributes of CUIs, such as negation, uncertainty, semantic group, and experiencer. We used negation and uncertainty attributes in our submitted system. In particular, we experimentally assigned weights to different 'values' of CUI attribute, as illustrated in Table 2. Those weights embody the importance of specific attribute values of CUIs in terms of ranking documents. Note that Table 2 corresponds to the weight assignment when there are no negated CUIs in the original query and the status of each query CUI is 'confirmed'.

Table 2. CUI attributes.

Attribute	Value	Description	Weight
Negation	pos	no negation	0.95
	neg	negated CUI	0.05
Uncertainty	sta0	confirmed	0.5
	sta1	history of	0.05
	sta2	family history of	0.1
	sta3	probable	0.35

Thus, an Indri query using CUI attributes will look like:

```
#weight(0.475 C0225386.sta0,pos, 0.0475 C0225386.sta1,pos,
0.095 C0225386.sta2,pos,  0.3325 C0225386.sta3,pos,
0.025 C0225386.sta0,neg,  0.0025 C0225386.sta1,neg,
0.005 C0225386.sta2,neg,  0.0175 C0225386.sta3,neg),
```

where we applied the attributes values as field restrictions in Indri, e.g., only negated C0225386 with 'probable' status in a document can match the Indri query term 'C0225386.sta3,neg', and consequently contribute with a weight value of 0.0175 to the relevance score of that document. Note that we used only one CUI in the above example for a simple demonstration. Longer CUI attribute queries can be formulated in a similar way.

3 Experiments and Results

We submitted 7 runs based on different settings of our retrieval system. For all the runs, we experimentally set the Dirichlet smoothing parameter to 2500. In this section, we describe each run in more detail, and present the corresponding evaluation results.

3.1 Submitted Runs

Table 3 illustrate the setting of each submitted run.

Table 3. Feature settings and evaluation scores for the submitted runs. 'X' responds to a selected feature.

	Feature Setting							Evaluation Measure	
RunNo.	MRM	MRF	MeSH	CUI	ATTR	DS_CUI	DS_CUI_ATTR	P10	NDCG@10
1								0.4720	0.4408
2	X	X		X		X		0.5180	0.4665
3	X	X	X	X		X		0.4880	0.4584
4	X	X	X	X	X	X	X	0.4820	0.4637
5	X	X	X					0.5040	0.4618
6	X	X	X	X				0.4940	0.4579
7	X	X	X	X	X			0.4700	0.4332

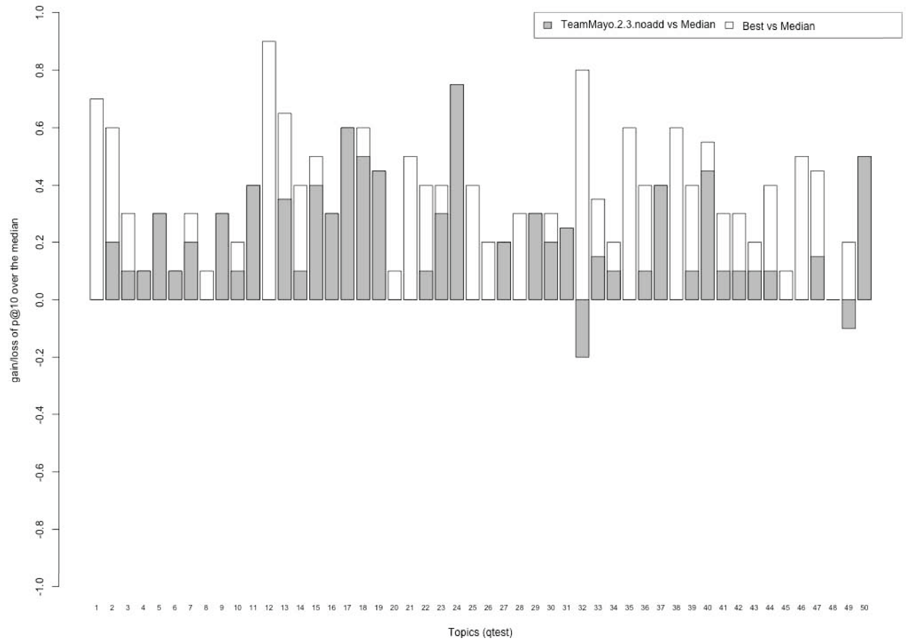
Run 1: This is our baseline run which exploited the query likelihood model (Eq. 1) but no other techniques or resources.

Run 2: In the text space, we obtained an initial ranking by using a linearly combined MRF and MRM model. We implemented this combined model in Indri by using another '#weight()' phrase, and trained the mixing weights based on the sample test collections provided by ShAReCLEF eHealth. Then, we re-ranked the top 1000 retrieved documents in the concept space by incorporating information in discharge summaries (see Sect. 2.3).

Run 3: On top of the setting for run 2, we further incorporated MeSH-based expansion (see Sect 2.2) for obtaining the initial ranking in the text space. Tests the third research question (usefulness of discharge summaries) when compared to run #6.

Run 4: On top of the setting for run 3, we further applied the CUI attributes (see Sect 2.3) for re-ranking documents in the concept space. Tests the third research question (usefulness of discharge summaries) when compared to run #7.

Fig. 2. Comparison of run 2 with CLEF median on metric P@10.



Run 5: We used the same setting as run 3 for the text-based retrieval, i.e., we leveraged MRF, MRM, and MeSH-based query expansion. However, we skipped the re-ranking process and used the initial ranking (i.e, rank list 1 in Fig. 1) for submission. Tests the first research question (robustness of IR methods from medical records) when compared to the baseline.

Run 6: On top of the setting for run 5, we further re-ranked the top 1000 retrieved documents in the concept space by using CUIs only (see Sect. 2.3). Tests the second research question (usefulness of NLP methods) when compared to run #5.

Run 7: On top of the setting for run 6, we further exploited CUI attributes (see Sect. 2.3) for documents re-ranking. Tests the second research question (usefulness of NLP methods) when compared to run #5.

3.2 Results and Discussion

Table 3 shows the official evaluation results for our submitted runs on the primary evaluation metrics P@10 and NDCG@10. Run 2 obtained the best overall performance among the seven runs. We further compare the P@10 performance of run 2 with the medians of all submitted runs of CLEF. As we can see in Fig. 2, for 14 out of the 50 test queries, run 2 obtained the best P@10 scores. Run 2 is also at or above the median for all but two of the queries (Topic 32 and Topic 49).

In error analysis, we found that the top 10 retrieved for Topic 32 (which contains a single query term ‘sob’, i.e., shortness of breath) are all non-English documents which are automatically considered as irrelevant according to the task guideline. This points to possible improvement through modifying the preprocessing algorithm, including a language filter. Our system performed reasonably on Topic 49, but other systems in the Task 3 pool outperformed it.

The comparison between the pairs (run 3 vs. 4, and run 6 vs. 7) suggests that the CUI attributes hurt the performance. This might be because due to the heuristic attribute weight settings (see Table 2), an issue due to the limited training data (only 5 sample queries were available for training).

The comparison between runs 3 and 6 and between 4 and 7 are insignificantly pessimistic about incorporating information in the discharge summaries into our retrieval model.

Furthermore, the performance of run 5 is comparable to run 2, suggesting that the text-based retrieval model alone is quite competitive. It is unclear whether the improved performance in run 2 is due to the absence of the MeSH terms, or the presence of the concept space.

4 Conclusion

We have shown that the text-based retrieval leveraging the well-tested MRF and MRM presented relatively strong performance for this CLEF retrieval task. The best-performing system also included NLP results and query expansion through discharge summaries. We have also found, however, that named entity attributes should not be used with untested heuristics. Future work includes a more principled inclusion of named entity attributes, different preprocessing of web text, and cross-validation to determine the stability of the results.

5 Acknowledgements

The work was supported by ABI: 0845523 from United States National Science Foundation, R01LM009959 from United States National Institute of Health. The challenge was organized by the Shared Annotated Resources (ShARe) project, funded by the United States National Institutes of Health with grant number R01GM090187.

References

1. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Mowery, D., Leveling, J., Goeuriot, L., Kelly, L., Martinez, D., Zuccon, G.: Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: CLEF 2013. Lecture Notes in Computer Science (LNCS), Springer (2013)
2. Voorhees, E.M., Tong, R.M.: DRAFT: Overview of the TREC 2011 medical records track. In: TREC. (2011)
3. 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. (2005)
4. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of the third ACM international conference on Web search and data mining, ACM (2010) 441–450
5. Hersh, W.: Information Retrieval: A Health and Biomedical Perspective. Third edn. Health Informatics. Springer (2009)
6. Liu, H., Wu, S.T., Li, D., Jonnalagadda, S., Sohn, S., Wagholikar, K., Haug, P.J., Huff, S.M., Chute, C.G.: Towards a semantic lexicon for clinical natural language processing. In: AMIA Annual Symposium Proceedings. Volume 2012., American Medical Informatics Association (2012) 568
7. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM (1998) 275–281
8. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics (1996) 310–318
9. 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, ACM (2001) 334–342
10. 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, ACM (2005) 472–479
11. Metzler, D., Croft, W.B.: Linear feature-based models for information retrieval. Information Retrieval **10** (June 2007) 257–274
12. Diaz, F., Metzler, D.: Improving the estimation of relevance models using large external corpora. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2006) 154–161
13. Zhu, D., Carterette, B.: Exploring evidence aggregation methods and external expansion sources for medical record search. In: Proceedings of The 21th Text REtrieval Conference. (2012)
14. Zhu, D., Carterette, B.: Combining multi-level evidence for medical record retrieval. In: Proceedings of the 2012 International Workshop on Smart Health and Wellbeing, Maui, USA, ACM (October 2012) 49–56
15. Hersh, W.R., Cohen, A.M., Ruslen, L., Roberts, P.M.: TREC 2007 genomics track overview. In: TREC. (2007)
16. Wu, S., Liu, H.: Semantic Characteristics of NLP-extracted Concepts in Clinical Notes vs. Biomedical Literature. In: Proceedings of AMIA 2011. (2011)

17. Wu, S., Liu, H., Li, D., Tao, C., Musen, M., Chute, C., Shah, N.: UMLS Term Occurrences in Clinical Notes: A Large-scale Corpus Analysis. In: Proceedings of the AMIA Joint Summit on Clinical Research Informatics. (2012)
18. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association **18**(5) (2011) 552–556
19. Zhu, D., Carterette, B.: Improving health records search using multiple query expansion collections. In: 2012 IEEE International Conference on Bioinformatics and Biomedicine, Philadelphia, USA (October 2012) 1–7
20. Aronson, A.R., Lang, F.M.: An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association (JAMIA) **17**(3) (2010) 229–36