ETH Zurich at TREC Clinical Decision Support 2016

Simon Greuter, Philip Junker, Lorenz Kuhn, Felix Mance,
Virgile Mermet, Angela Rellstab and Carsten Eickhoff

Department of Computer Science
ETH Zurich, Switzerland
carsten.eickhoff@inf.ethz.ch

Abstract

This paper describes ETH Zurich's submission to the TREC 2016 Clinical Decision Support (CDS) track. In three successive stages, we apply query expansion based on literal as well as semantic term matches, rank documents in a negation-aware manner and, finally, re-rank them based on clinical intent types as well as semantic and conceptual affinity to the medical case in question. Empirical results show that the proposed method can distill patient representations from raw clinical notes that result in a retrieval performance superior to that of manually constructed case descriptions.

1 Introduction

The volume of annually published scholarly medical articles has been growing rapidly in recent years. Statistics report a growth of the MedLine directory by as much as 1 Million new citations per year. While this considerable amount of scientific research holds a rich and ever increasing well of knowledge, its sheer scale makes it intractable for manual inspection and mandates the development of dedicated automatic retrieval facilities.

In this paper, we present a modular patient-centric information retrieval system based on a pipeline of individual query and document processing steps. Most notably, our system provides functionality for query expansion, document retrieval as well as a number of re-ranking methods. Starting from noisy natural language health records in the form of clinical notes, we apply rigorous filtering to reduce the effects of surface-form variance between notes and articles.

Our query expansion scheme jointly combines evidence from 4 sources of information: (1) human expertise in the form of explicitly assigned keywords, (2) textual statistics of highly salient tf-idf terms, (3) external ontological MeSH information (4) semantics-preserving neural word embeddings. After an initial retrieval run pseudo-relevant documents are analyzed and evidence for the usefulness of expansion candidates is aggregated from our various sources, resulting in a significantly more powerful and robust representation of the clinical information need. Using this augmented query representation, our retrieval model is centrally based on a BM25 variant that is capable of detecting natural language negations. As such, negated terms and their respective scopes can be treated differently from findings recorded in positive modality. The model allows us to carefully distinguish between confirmed and refuted facts in notes as well as scientific articles, leading to a significantly more intricate relevance estimation than mere keyword matching could achieve.

While the previous steps aimed at a reliable representation of topic and document subject matter, each clinical case is accompanied by a specific context that may be independent of the core medical findings. To this end, we apply multiple types of re-ranking, each maximizing the likelihood of retrieved documents given the topic's intent type (diagnosis, test, or treatment), its latent semantic focus as well as conceptual similarities. Experiments on historic editions of the TREC CDS track led

25th Text Retrieval Conference (TREC) Notebook 2016.

Table 1: Parameters for term-wise query expansion.

	keywords	tf-idf	MeSH	α	β	n
summary	$k = 100$	$k = 20$	$k = 50$	0.5	0.5	60
description	$k = 100$	$k = 20$	$k = 50$	0.4	0.4	100
note	$k = 100$	$k = 20$	$k = 50$	0.4	0.4	80

to very encouraging results, especially when working with the generally harder-to-process long and noisy descriptions. Given that the newly introduced notes showcase even higher degrees of noise and initial retrieval difficulty, we believe that such functionality is crucial for attaining satisfactory retrieval clinical decision support performance.

2 Methodology

The submission relies on three successive steps, (1) query expansion enriches raw patient records with additional synonyms and semantically related terms. (2) A negation-aware ranking model computes document relevance scores. (3) Finally, several re-ranking components modify these raw scores to give the final ranking. The following sections describe the respective components in detail.

2.1 Term-wise Query Expansion

This expansion method uses terms proposed by pseudo relevance feedback and is composed of three components that use different aspects of each pseudo-relevant document. We begin by lowercasing the query and remove stop words. In addition to that, we filter the noisier descriptions and notes to retain only nouns and verbs. The presented method expands an initial query q by a set of terms E.

$$q' = q + E \tag{1}$$

In a first retrieval run using query q, we obtain a ranked list of the k most relevant documents $D_{k,q}$. All documents $d_i \in D_{k,q}$ are sorted in descending order of relevance. The set of newly added terms E is given by the n most important terms according to $P_C(t, D_{k,q})$ in $D_{k,q}$.

$$E = argmax_{t,n} P_C(t, D_{k,q}) \tag{2}$$

A term's importance is supported by three sources of evidence. (1) The document's keyword meta-information field. The importance score $P_W(t, D_{k,q})$ represents the number of occurrences of keyword t in $D_{k,q}$. (2) High-ranking tf-idf terms are extracted from the whole article. $P_S(t, D_{k,q})$ indicates the number of times that term t has a high tf-idf ranking in $D_{k,q}$. (3) MeSH concepts are extracted from article abstracts. $P_M(t, D_{k,q})$ counts the occurrence of individual MeSH terms. The overall importance $P_C(t, D_{k,q})$ is calculated on the basis of the normalized importance scores obtained from the three previously discussed sources of evidence and weighted using factors α, β and γ, which we require to sum up to one.

$$P_C(t, D_{k,q}) = \alpha \cdot \|P_W(t, D_{k,q})\| + \beta \cdot \|P_S(t, D_{k,q})\| + \gamma \cdot \|P_M(t, D_{k,q})\| \tag{3}$$

The most reliably performing parameters are shown in Table 1.

2.2 Semantic Query Expansion

To go beyond literal term matches and expand queries with semantically related terms, we employ Google's word2vec[1]. The skip-gram model is first used to train distributed word vector representations on the collection of all preprocessed documents (topics t_i and articles a_i of 2016). The so learned vocabulary is further refined with a smaller learning rate into per-topic vocabularies V_{t_i} and V_{qrels}.

V_{t_i} is trained on the concatenated phrases of topic t_i and articles retrieved for this topic by a standard BM25 retrieval run.

[1] https://code.google.com/archive/p/word2vec/

V_{qrels} is learned by a modified version of word2vec. Given a sentence $s = w_0, w_1, \ldots, w_{k-1}, w_k$, the standard skip-gram task is to predict a local context around each word, e.g., given the skip-word w_i, the neural net is judged on how well it predicts words

$$w_{i-c}, w_{i-c+1}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+c-1}, w_{i+c}$$

for a context size of $2c$. To train V_{qrels}, we replace words from local context by random words from an article we are confident to be relevant to the topic w_i belongs to. As ground-truth collection C we use the topic/article pairs from qrels2014 and qrels2015, as well as the top ranked results for 2016 retrieved by a standard BM25 baseline.

For each original word w_o in a query, the global V_{qrels} and the per-topic embedding V_{t_i} are used to find its k neighbours maximizing the cosine similarity in each embedding.

To achieve a focus on medical expansion terms, those of the $2k$ neighbours that appear less frequently than a given threshold ω_0 in a general news corpus are considered as expansion candidates e_i. Additionally, candidates stemming from V_{t_i} are discarded if they are not used as a keyword in at least one article $a_i \in C$.

From the remaining candidates, expansion terms are picked in a round-robin fashion up to a maximum ratio of original words w_o vs. expansion terms e_i of $\frac{count(e_i)}{count(w_o)} \leq \omega_1$.

2.3 Negation-aware Ranking

Investigations by Kuhn and Eickhoff [7] on historic TREC CDS data suggest that, if not taken into account appropriately, the presence of negations in medical case records can have a significant negative impact on retrieval performance. Motivated by these findings, our retrieval model accounts for sentences in queries and documents, in which the author explicitly notes the absence of some symptom or condition. *"She denies smoking, diabetes, hypercholesterolemia, or a family history of heart disease."*, for instance.

Following Limsopatham *et al.* [8], our approach is based on introducing a new term representation for terms appearing in a negated context, *i.e.*, *"no diabetes"* is converted to *"[nx]diabetes"*. The intuition behind this approach is to match the contexts (positive or negative) in which a given term appears in both the query and the document.

Using NegEx [2], we tag all negated terms in the document collection and the queries in the above manner. Furthermore, the untagged versions of negated terms are added to the queries, with reduced term weights, to extend the coverage of our queries.

After pre-processing the data in this fashion, we compute relevance scores according to the standard Okapi BM25 retrieval model [10].

$$S(Q, D) = S(Q_{tagged}, D_{tagged}) + \beta \times S(Q_{neg}, D_{tagged})$$

where Q_{tagged} and D_{tagged} are the negation-tagged versions of the query and the document respectively. Q_{neg} are the negated terms from the query in untagged form.

2.4 Re-ranking

To account for variability in the concrete choice of words between documents and queries, we re-rank the literal matching scores described previously based on estimates of topical affinity.

We use a latent dirichlet allocation (LDA) [1] model with 50 topics and a vocabulary of 150’000 terms. The model is trained by iterating 25 times over an in-domain training-set containing 114’000 bio-medical journal articles. To extract the in-domain training set from the provided TREC corpus, we select the 4000 top-ranking documents of each query according to BM25.

The LDA model takes a text document in bag-of-words representation as input and returns a multinomial distribution over latent topics. The topic affinity score, describing the similarity between a query and a document, is obtained in the form of the Jensen-Shannon (JS) divergence between topic distributions of query and document as shown in Equation 4, where Q_T and D_T are the topic distributions of query Q and document D.

$$S_{\text{topic-model}}(Q_T, D_T) = 1 - JSD(Q_T, D_T) \tag{4}$$

The Jensen-Shannon (JS) divergence is a smoothed and symmetrized version of the Kullback-Leibler (KL) Divergence which measures how bad the probability distribution P is at modelling S and vice versa. The JS-divergence is bounded by $[0, 1]$ and defined in the following way:

$$JSD(P, S) = \frac{1}{2} * KL(P, M) + \frac{1}{2} * KL(S, M) \tag{5}$$

Where M is responsible for the symmetry in the JS-Divergence and defined as follows:

$$M = \frac{1}{2} * (P + S) \tag{6}$$

$$KL(P, S) = \sum_i P(i) * log_2(\frac{P(i)}{S(i)}) \tag{7}$$

In addition to topical re-ranking we further measure the likelihood of document D satisfying the clinical intent type (diagnosis, treatment, or test) specified in topic T. To this end, we employ an idea similar to the classification and fusion approaches of [3, 11, 4]. For each topic-document pair (T, D), we compute a classifier score $S_{\text{classifier}}(T_{\text{type}}, D)$ which measures to which degree document D matches the intent type of topic T.

The classifier score is given by:

$$S_{\text{classifier}}(T_{\text{type}}, D) = \frac{S_{\text{ML-classifier}}(T_{\text{type}}, D) + S_{\text{Keyword-Counter}}(T_{\text{type}}, D)}{2} \tag{8}$$

The first component of Equation 8 is computed using linear SVMs with a *squared* loss function (using the implementation from `scikit-learn` [9]). For each of the three intent types — *diagnosis*, *test*, *treatment* — we construct a binary linear SVM classifier; topics 1—10 use the *diagnosis* classifier, topics 11—20 use the *test* classifier, and topics 21—30 use the *treatment* classifier. The classifier is applied to the top 1000 results of each topic.

The positive training samples of the classsifiers are PMC articles retrieved by the following PMC search queries:

- for the *diagnosis* and *test* classifiers, we use the following PMC query: `open access[filter] AND diagnosis[MeSH Major Topic];`
- for the *treatment* classifier, we use the following PMC query: `open access[filter] AND therapeutics[MeSH Major Topic];`

The negative training samples are the rest of the MeSH-indexed documents in the collection.

The second component of Equation 8, the *Keyword-Counter*, measures for each document the frequency of certain keywords related to the topic intent type:

- for *diagnosis*, frequency of words that stem to *diag*;
- for *test*, frequency of words that stem to *diag* or *test*;
- for *treatment*, frequency of words that stem to *treat*.

The classifier score and the Keyword-Counter score are then normalized and fused according to Equation 8, to produce the final classifier score. This score is then fused with the original ranking scores, as well as the previously described topical affinity scores and the query-document cosine similarity in doc2vec space using linear combination according to Equation 9:

$$\begin{aligned} S_{\text{final}}(T, D) = {} & \alpha \cdot S_{\text{relevance}}(T, D) \\ & + \beta \cdot S_{\text{classifier}}(T_{\text{type}}, D) \\ & + \gamma \cdot S_{\text{topic-model}}(T, D) \\ & + \delta \cdot S_{\text{doc2vec-model}}(T, D) \end{aligned} \tag{9}$$

Table 2: Linear combination weights for Equation 9.

	α	β	γ	δ
Diagnosis	0.75	0.125	0.11875	0.00625
Test	0.85	0.015	0.1215	0.0135
Treatment	0.8	0.1	0.095	0.005

where $\alpha, \beta, \gamma, \delta$ sum up to 1.

For TREC 2016, we choose the weights that maximize the mean BM25 scores of the TREC 2014 and 2015 topics. These weights are listed in Table 2:

After the TREC 2016 submission deadline, we further improved the ML-classifier in several ways:

- we used the confidence values of the decision function as the ML-classifier scores, instead of the predicted class labels (0 or 1);
- we used the *error-rate* loss function (from `SVM-perf`[6]), instead of the *squared* loss function;
- we used the Clinical Hedges Database [5] for training, instead of documents retrieved by PMC queries; the CHD contains 1000 positive diagnosis documents and 8000 positive treatment documents.

In the following section, we will discuss both the official submitted results as well as the performance of the more recent, modified re-ranking scheme.

3 Results

We submitted five official TREC CDS 2016 runs for evaluation, addressing all three query types (summary, description, note). During previous experiments, negation-aware ranking and query expansion consistently improved retrieval performance. As a consequence, we apply the techniques described in Sections 2.1 – 2.3 in all five runs. The various re-ranking methods previously showed to be highly parameter dependent and could occasionally decrease performance. To account for this fact, we discuss scores both before and after re-ranking. Table 3 lists the final performance of each run in terms of precision at 10 retrieved documents. Official runs that were submitted to TREC 2016 are highlighted with an asterisk.

Table 3: Recomputed experiments results (P@10).

	Expanded	**Re-ranked Submitted**	**Re-ranked Improved**
Summaries	30.33*	30.67*	32.67
Descriptions	23.33	22.67*	26.33
Notes	25.67*	25.67*	29.33

While the preliminary submitted re-ranking showed no noteworthy influence on result quality, the improved version has a consistently positive effect. Similar to the results of previous years, summaries tended to make better queries than the longer descriptions. Interestingly, however, our pre-processing was able to distill considerable amounts of information from raw clinical notes, making them more effective queries than the manually produced descriptions.

4 Conclusion

In this paper, we presented a three-stage processing pipeline for clinical notes including query expansion, negation-aware ranking and finally, a re-ranking step. We obtained a solid overall performance and were able to achieve, on the basis of raw notes, a retrieval performance superior to that of manually created case descriptions.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(933-1022), 2003.

[2] Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301 – 310, 2001.

[3] Sungbin Choi and Jinwook Choi. Snumedinfo at TREC cds track 2014: Medical case-based retrieval task. 2014.

[4] Eva D'hondt, Brigitte Grau, and Pierre Zweigenbaum. Limsi@ 2015 clinical decision support track.

[5] Brian Haynes, Ann McKibbon, Nancy Wilczynski, Stephen Walter, and Stephen Werre. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ*, 330(7501):1179, 2005.

[6] Thorsten Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 377–384. ACM Press, 2005.

[7] Lorenz Kuhn and Carsten Eickhoff. Implicit negative feedback in clinical information retrieval. *Medical Information Retrieval Workshop (MedIR), Pisa, Italy*, 2016.

[8] Nut Limsopatham, Craig Macdonald, Richard McCreadie, and Iadh Ounis. Exploiting term dependence while handling negation in medical search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1065–1066. ACM, 2012.

[9] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[10] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, 109:109, 1995.

[11] Ronghui You, Yuanjie Zhou, Shengwen Peng, and Shanfeng Zhu. Fdumedsearch at TREC 2015 clinical decision support track.