Implicit negative feedback in clinical information retrieval

Lorenz Kuhn, Carsten Eickhoff

Department of Computer Science, ETH Zurich, Switzerland

Summary

In this paper, we reflect on ways to improve medical information retrieval accuracy by drawing implicit negative feedback from negated information in noisy natural language search queries. We begin by studying the extent to which negations occur in clinical texts and quantifying their detrimental effect on retrieval performance. Subsequently, we present approaches to query reformulation and ranking that remedy these shortcomings by resolving natural language negations. Our experimental results are based on data collected in the course of the TREC Clinical Decision Support Track and show consistent improvements compared to state-of-the-art methods. For queries that make excessive use of negations, we were able to achieve up to 300% relative improvement in early precision.

Key words: medical information retrieval, negation detection, natural language processing, negative feedback

Introduction

Whether it is finding an appropriate test, making a diagnosis or suggesting a treatment, clinical decision making is a challenging task. Finding relevant information for the wide variety of health problems physicians encounter on a day-to-day basis is difficult and time-consuming. As a result of the exponential increase in the amount of research articles published annually, manually identifying the most important and relevant texts has become a hard task.

State-of-the-art retrieval models applied to clinical decision-support settings rely on full-text indices of biomedical literature and use the textual content of the patient record to construct queries. These models were designed with keyword search interaction in mind, but medical case narratives are maintained in natural language, resulting in significantly longer queries than those we are used to in Web search settings. In our case study, the average query length after removing stop words was 55.7 words.

Besides their mere length, negations represent a particularly challenging aspect of natural language queries. Consider the following example, taken from Topic 1 of the TREC 2014 Clinical Decision Support Track: “She denies smoking, diabetes, hypercholesterolemia, or a family history of heart disease.” The clinical practitioner encoded explicit knowledge of the absence or invalidity of a range of conditions or findings but our term-based retrieval model readily uses the entire negated passage as query terms. This inappropriate use of the carefully curated clinical narrative results in measurable detriments in retrieval performance. We quantified this effect by comparing two sets of TREC 2014 case reports: those containing no negations D+ (14 reports), and those containing at least some negated information D− (16 reports). We found a clear negative impact of the presence of negated terms on the retrieval results. Both normalised discounted cumulative gain (nDCG; 25% improvement) and P@10 (9.6% improvement) were significantly higher for D+ than D−.

This observation is not just limited to small academic collections such as the TREC corpus, but also holds in real-world clinical environments. Chapman et al. [3] found between 39% and 83% of all clinical observations to be described in a negated form. In this paper, we empirically compare state-of-the-art query filtering techniques, as well as novel query-adaptive retrieval models, that actively use negated terms as negative relevance feedback. Our investigation was based on corpora and relevance judgements of the TREC 2014 Clinical Decision Support Track and highlights the merit of the proposed method.

Background

This study built on previous findings from both automatic negation detection in natural language processing and negative relevance feedback for retrieval models. The following paragraphs summarise the most relevant developments in both fields.

Rokach et al. [7] provided an extensive overview of negation recognition methods for medical narrative reports. The previous work could be categorised into knowledge engineering and machine learning-based approaches. We will discuss one representative example per category. Chapman et al. [4] proposed NegEx, a regular expression-based algorithm to detect negated findings in radiology reports. Testing this algorithm on 1235 findings and diseases in 1000 sentences taken from discharge summaries, NegEx
achieved a specificity of 94.5% and a sensitivity of 77.8%. As an example of machine-learned negation detectors, Agrawal et al. [1] presented a conditional random field model, designed to detect negation cues and their respective scopes. The model was trained on the publicly available BioScope corpus [8]. This approach outperformed NegEx with F1 scores of 98% for detecting cues and 95% for detecting scopes.

The field of information retrieval has long-standing experience in using feedback of (pseudo) relevance in the retrieval process [6]. However, the use of explicit non-relevance information has been shown to be more difficult to incorporate. Wang et al. [9] investigated different methods to improve retrieval accuracy for difficult search queries, using negative feedback. Their work covered both language and vector-space models, as well as a number of heuristics for negative feedback. In the Score Combination strategy, a positive query representation $Q$ and a negative query representation $Q_{neg}$ were maintained separately. The scores for a given document were computed for both query representations and then combined for the final result.

Previous approaches to using negations in medical information retrieval have focused on removing negated terms completely. Averbuch et al. [2] were able to improve F scores by 8.28% on average by removing negated UMLS terms from queries. Even though this approach has been shown to improve retrieval results, a lot of information is lost altogether by filtering of negated terms from the query. In the following, we propose a way of explicitly using such negated information to improve retrieval performance.

Case study

General setup

Our empirical investigation is based on the TREC 2014 Clinical Decision Support (CDS) track document collection. The corpus consists of an open-access subset of PubMed Central, an online repository of biomedical literature, as well as a number of artificial, idealised medical case reports, created by experts at the US National Library of Medicine. In accordance with the track’s guidelines, our retrieval experiments used the full text narrative of these reports as queries.

The document collection was indexed using Apache Lucene, with default settings. After the inspection of a broad method and parameter sweep, we relied on an Okapi BM25 retrieval model [5] which delivered consistently strong results.

For our queries, we extracted the description of the provided topics. We applied lower casing and removed stop words. In the following, we will utilise three different versions of queries:

- The full description ($Q_{full}$)
- The description, from which all negated sub-sentences were removed ($Q_{pos}$)
- The negated sub-sentences ($Q_{neg}$)

As a proof of concept, negations and their scopes were initially annotated manually. Empirical comparison with NegEx [4] showed only negligible differences that did not have a noticeable effect on retrieval performance.

Methods

Filtering

The traditional way of addressing negations in natural language queries, as investigated by [2], simply removes negated sub-sentences from the query. The score for a document $D$ and query $Q$ is computed as:

$$S(Q, D) = S(Q_{pos}, D)$$

where $S(Q, D)$ is the BM25 score of document $D$ for query $Q$.

Score combination

Although the filtering approach to negation handling has been shown to perform well in practice, intuition mandates that making explicit use of the information contained in the negation should be beneficial. To this end, we relied on the score combination method of Wang et al. [9], which computes the relevance score for query $Q$ and document $D$ as:

$$S_{combined}(Q, D) = S(Q_{full}, D) - \beta \cdot S(Q_{neg}, D)$$

We adapted this method to our needs by constructing $Q_{neg}$ from the negated query terms, instead of using negative document examples. We denoted the number of terms in the current query as $n_{full}$ and the number of negative terms as $n_{neg}$. To avoid assigning too much weight to negative terms if they occurred infrequently, we set $\beta$ in the following, empirically determined manner:

$$\beta = 2.5 \cdot (n_{neg}/n_{full}), \text{ if } (n_{neg}/n_{full}) > 0.25, \text{ else } \beta = 0$$

As the number and extent of negated phrases among the provided queries was relatively low (on average 3.97 words per 56-term query), the impact of both negation filtering

<table>
<thead>
<tr>
<th>Table 1: Comparison of methods, all Topics.</th>
</tr>
</thead>
<tbody>
<tr>
<td>P@10</td>
</tr>
<tr>
<td>Baseline</td>
</tr>
<tr>
<td>Negation filtering</td>
</tr>
<tr>
<td>Score combination</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Comparison of methods, Topic 1.</th>
</tr>
</thead>
<tbody>
<tr>
<td>P@10</td>
</tr>
<tr>
<td>Baseline</td>
</tr>
<tr>
<td>Negation filtering</td>
</tr>
<tr>
<td>Score combination</td>
</tr>
</tbody>
</table>
and score combination is limited. Nevertheless, our method not only consistently outperformed the baseline, but also improved on the established negation filtering strategy in all considered metrics (see table 1).

When considering those topics that contain significant amounts of negated information (e.g., Topic 1 with 30% of all terms occurring in a negated context), both of the methods greatly improve P@10. Where negation filtering achieves a 200% relative improvement, our proposed score combination method resulted in an even more pronounced gain of up to 300% relative improvement (see table 2). Furthermore, whereas negation filtering detrimentally affected nDCG and infAP, score combination outperformed the baseline in both of those aspects by leaving queries with limited degrees of negated information unaltered.

Limitations

Clearly, the interpretation of the results presented here is limited by the small sample size as well as the relative brevity of case reports. Real-world medical case narratives often span several pages or volumes as the patient history unfolds across years of treatment. The observed benefit of using negative feedback methods is difficult to assess on artificially generated corpora and may require investigation of more sizable real-world collections.

Conclusion

Making use of negative information is critical for retrieving documents in clinical contexts. In this paper, we have laid out how automatic negation detection output can be utilised by actively discounting documents containing negated query terms. Our case study indicates that this approach is more promising than ad-hoc removal of negated terms. Empirical results show a small but consistent improvement across all queries, as well as a larger quality gain for those topics that contain negations more frequently. There are several interesting research questions that we aim to address in the future.

1. This work studied a small academic sample of carefully curated artificial case reports. In the future, it will be mandatory to investigate the generalisability of our findings to real-world collections of considerable size.

2. Similarly, we aim to investigate the effect of going beyond the currently studied short and artificial patient records towards longer clinical narratives.

3. Finally, in the future, adaptive choices of \( \beta \) should account for the actual importance of negated terms and not just their relative length.

Disclosure statement

No potential conflict of interest relevant to this article was reported.

References


Correspondence:

Dr. Carsten Eickhoff
Departement Informatik der ETH Zürich
Universitätstrasse 6
CH-8092 Zürich
ecarste[at]inf.ethz.ch