Understanding Negation and Family History to Improve Clinical Information Retrieval

Bevan Koopman  
Australian e-Health Research Centre, CSIRO  
Brisbane, Australia  
bevan.koopman@csiro.au

Guido Zucon  
Queensland University of Technology  
Brisbane, Australia  
g.zuccon@qut.edu.au

ABSTRACT
We present a study to understand the effect that negated terms (e.g., “no fever”) and family history (e.g., “family history of diabetes”) have on searching clinical records. Our analysis is aimed at devising the most effective means of handling negation and family history. In doing so, we explicitly represent a clinical record according to its different content types: negated, family history and normal content; the retrieval model weights each of these separately. Empirical evaluation shows that overall the presence of negation harms retrieval effectiveness while family history has little effect. We show negation is best handled by weighting negated content (rather than the common practise of removing or replacing it). However, we also show that many queries benefit from the inclusion of negated content and that negation is optimally handled on a per-query basis. Additional evaluation shows that adaptive handling of negated family history content can have significant benefits.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]
General Terms: Measurement, Experimentation.

1. INTRODUCTION
Negation and reference to family history are two unique characteristics of clinical records that affect natural language processing of clinical text [3]. Commonly mentioned conditions in a patient record (e.g., “fever” or “fracture”) often appear in negated form (e.g., “denies fever” or “no fracture”) [3]. Previous research has largely focused on identifying the negated portions of text, or reference to family history content (e.g., “family history of heart disease”) [3, 1]. From an information retrieval (IR) perspective, previous studies have considered how negation may adversely affect retrieval [5, 6]. For example, when searching patient records using the query “patients with heart murmur”, the retrieval system might return a large number of irrelevant documents containing “no heart murmur”. Traditional keyword-matching IR systems denote the presence of the query terms in a document as an indicator of relevance. Empirical analysis of the effect of negation on IR system effectiveness shows mixed results: Koopman et al. [5] found IR term-weighting methods naturally accounted for negation, while Limspopatham et al. [6] developed a technique that showed handling negation improved effectiveness.

In this paper, we provide a specific analysis of how and why negation affects retrieval effectiveness. In doing so, we uncover why previous studies of negation in IR produced differing results. In addition to negation, we also consider how the reference to family history also influences IR effectiveness. We explicitly represent, within the language modeling framework, a clinical record according to its different content: negated, family history and normal (i.e., all other normal content); the importance of each of these content types is then weighted separately. Relevance of a particular document to a query is estimated based on the mix of the three content types within the document.

An evaluation using the TREC Medical Records Track shows that handling negation does improve retrieval performance. However, further analysis revealed that many queries benefit from the inclusion of negated content. An outcome of this finding is that the common approach [9, 8] of removing or replacing negated content from the document representation is sub-optimal; rather, negated content should be weighted separately, ideally on a per-query basis. The significant potential benefits of adaptive per-query handling of negated, family history and normal content are presented in further retrieval experiments.

2. RELATED WORK
Accounting for negated terms in clinical text has been an important topic in health informatics, with much of the focus being within the computational linguistics and natural language processing (NLP) fields. The main focus of these efforts is on negation detection and negated scope detection. Chapman et al. [2] developed NegEx, an algorithm which is effective in determining negated findings or diseases from clinical text. NegEx has become a common tool for identifying negated content; the tool was extended as the ConText algorithm, which in addition to negation also identifies hypothetical, or historical references in clinical text [4]. ConText was also extended to identify references to family history. Previous studies reported the effectiveness of NegEx to be at least 90% F-measure [2, 4].

Less research has been performed on the effect of negation on searching clinical text. Previous studies in this area have mainly considered how the negated content of a document can be removed or separated prior to indexing the documents; the assumption being that the presence of negated content always harms retrieval effectiveness [1, 6]. This assumption was pervasive amongst teams participating in the TREC Medical Record Track: many participants dealt with negation by pre-processing clinical records with the NegEx algorithm to remove negated content [9, 8].
In this paper, we firstly empirically investigate the assumption that negation always harms retrieval performance. This is important to understand as previous studies differ in their findings on the effect of negation: Koopman et al. [5] found IR term-weighting methods naturally accounted for negation, while Limsoopatham et al. [6] developed a technique that showed penalising negation improved effectiveness. In addition to the effect of negation, we also consider the less studied effect of family history references on clinical IR.

The previous work described here, and the focus of this study, is on explicitly negated terms found in documents, which differs from other work concerned with negation in queries (e.g., the Boolean query “hypertension NOT obesity”). Dealing with negation in queries presents its own set of challenges but is out of the scope of this study.

3. RETRIEVAL MODEL

We model retrieval as a language modeling process, where the standard document representation is enhanced to handle three different types of content within the document: negated, family history and normal. To this aim, we separate these different contents such that a document D can be represented by $D = D_{nor} \cup D_{neg} \cup D_{fh}$.

Separating the different contents allows weighting each content type differently. For example, instead of removing negated content from the document, a negative weight can be assigned to the negated content when estimating the probability of a document being relevant to a query Q:

$$P(Q|\hat{D}) \approx P(Q|D_{nor}) - P(Q|D_{neg})$$ (1)

This approach of subtracting the score contribution of negated content is similar to Limsoopatham et al. [6], who reported improvement in retrieval effectiveness using this approach. However, we further enhance this by mixing the estimates of negated, family history and normal content:

$$P(Q|\hat{D}) \approx \lambda_{nor} P(Q|D_{nor}) + \lambda_{neg} P(Q|D_{neg}) + \lambda_{fh} P(Q|D_{fh})$$ (2)

where $\lambda_{nor, fh, neg}$ are the mixing parameters that control the weights for normal, negated and family history content respectively. Weights of normal and family history are bounded by $0 \leq \lambda_{nor, fh} \leq 1$, however negation weights are instead bounded by $-1 \leq \lambda_{neg} \leq 1$; this is done to handle negative weighting for negation (making it equivalent to Eq. 1). Eq. 2 explicitly scores a clinical record according to its different types of content; the effect of negation and family history can then be investigated by varying the mixing parameters $\lambda$.

The Indri toolkit\(^1\) was used to implement Eq. 2. For negation and family history detection we used the standard ConText algorithm [4]. Spans of text in a document are annotated with XML elements &lt;negated&gt; or &lt;history&gt;, all other content is annotated &lt;normal&gt; (spans may overlap). Documents are then indexed using Indri’s XML indexer, which stores the normal, negated and family history content in separate fields, thus providing the three representations of a document, $D_{nor}$, $D_{neg}$ and $D_{fh}$. For retrieval with mixed weights we used the Indri Query Language #wsum method to assign weights to specific fields, e.g., if a query text is Dementia, and the mixing parameters are $\lambda_{nor} = 1.0$, $\lambda_{neg} = -1.0$ and $\lambda_{fh} = 0.5$, we generate the query #wsum(1.0 dementia.normal -1.0 dementia.negated 0.5 dementia.history). This query can be interpreted as requesting medical records than mention dementia but not in a negated form; while, the mention of dementia in the family history is weighted half that of an affirmative mention in the normal content.

4. EMPIRICAL EVALUATION

Our experimental evaluation is conducted to answer the following research questions:

RQ1: What effect does negation have on overall retrieval effectiveness?
RQ2: What effect does family history have on overall retrieval effectiveness?
RQ3: Does negation always harm effectiveness (and therefore should we always exclude or negatively weight negated content)?
RQ4: Can retrieval effectiveness be significantly improved by finding an optimal mix of normal, negated and family history content?

The test collection used in our experiments was the TREC 2011 & 2012 Medical Records Track (MedTrack) [8, 9]. The unit of retrieval was a patient record rather than an individual report; thus, reports belonging to a single patient’s record were concatenated into a single document called a patient visit document.\(^2\) The resulting corpus contained 17,198 patient visit documents.

To evaluate RQ1 (the overall effect of negation), we adjust the weight of negated content while fixing the weights for normal and family history. This is done by varying $\lambda_{neg}$ in Eq. 2 from -1 to 1 in 0.1 increments. Negation removal equates to $\lambda_{neg} = 0$: this is the approach most systems at TREC Medtrack subscribe to, including the system that achieved the highest results (Udel) [8, 9]. Note that in our experiments we do not explicitly compare with TREC systems because they mix negation handling (if any) with other techniques and engineering solutions (such as discriminating between type of reports). Instead, we consider two baselines to handle negation: that with $\lambda_{neg} = 0$ (negation removal), which is used by most TREC system, and that with $\lambda_{neg} = -1$ (negation penalisation), which resembles the strategy by Limsoopatham et al. [6].

To evaluate RQ2 (the overall effect of family history) we adjust the weight of family history content while fixing the weights for normal and negated by varying $\lambda_{fh}$ in Eq. 2 from 0 to 1 in 0.1 increments.

To evaluate RQ3 (does negation always harm performance) we have a two-fold approach. Firstly, we investigate individual query performance for the best settings of $\lambda_{neg}$. Secondly, we perform an exploration of the full parameter space of $\lambda_{nor}$, $\lambda_{neg}$, $\lambda_{fh}$ to uncover how the weighting mix differs between queries and whether negation always harms

\(^1\)Lemur Project, http://www.lemurproject.org

\(^2\)This is a common practise among MedTrack participants [8, 9].
Table 1: TREC MedTrack retrieval results for negation removal, family history removal and both negation and family history removal. † = statistical significance (paired t-test, p < 0.01) over the baseline.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\lambda_{nor}$</th>
<th>$\lambda_{neg}$</th>
<th>$\lambda_{fh}$</th>
<th>bpref</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (include all)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.3644</td>
<td>0.4469</td>
</tr>
<tr>
<td>Negation removal</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.3811</td>
<td>0.4901</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(+5%)†</td>
<td>(+10%)†</td>
</tr>
<tr>
<td>Family history removal</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.3652</td>
<td>0.4519</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(+0.2%)</td>
<td>(+1.1%)</td>
</tr>
<tr>
<td>Negation &amp; family history removal</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.3813</td>
<td>0.4914</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(+5%)†</td>
<td>(+10%)†</td>
</tr>
</tbody>
</table>

Figure 1: The effect on retrieval effectiveness ($y$-axis) with different weightings ($x$-axis) of negation (left) and family history (right).

RQ1 & RQ2: What effect does negation and family history have on overall retrieval effectiveness?

Table 1 presents the retrieval results for negation removal, family history removal and both negation and family history removal. Removing negated content does indeed improve overall retrieval performance, especially in P@10. The greater improvement in P@10 shows that negation can more adversely affect the top-ranked results. The results confirm that negation has an adverse effect on retrieval effectiveness and that removal of negated content ($\lambda_{neg} = 0$) can improve the overall effectiveness. RQ2 considers the effect of family history content on retrieval effectiveness. In this regard, removing family history content did not affect effectiveness in a statistically significant way: Table 1 shows only minor changes in P@10 and bpref for family history removal.

The results in Table 1 report the effect of simply removing negation and family history; in our method, this corresponded to assigning a weight of 0 to $\lambda_{neg}$ and $\lambda_{fh}$. However, Eq. 2 also allows assigning different weights to the different types of content. The effect on retrieval for different weighting values is illustrated in Figure 1. The $x$-axis refers to different values of $\lambda_{neg}$ from -1.0 to 1.0 and $\lambda_{fh}$ from 0.0 to 1.0. In both cases, the weight for normal content is fixed, $\lambda_{nor} = 1.0$. The $y$-axis shows the retrieval effectiveness (bpref and P@10).

For negation weighting, Figure 1 shows that negative weights of $\lambda_{neg}$ are more effective (for both bpref and P@10) than negation removal ($\lambda_{neg} = 0$). This finding is consistent with Limsdopatham et al. [6] and supports their negative weighting method. It also highlights that common approaches of simply removing negated content [9, 8] are sub-optimal; a more appropriate method is to negatively weight such content. The optimal effectiveness differs depending on the evaluation metric — bpref is maximised with $\lambda_{neg} = -1.0$, while P@10 is maximised with $\lambda_{neg} = -0.7$. This difference again highlights that negation affects the top-ranked documents (measured by P@10) in a different way to the longer result set (measured by bpref).

In contrast to negation, weighting the contribution of family history content has little effect on effectiveness. However, it is important to note that removing the document portions that dealt with family history (i.e., $\lambda_{fh} = 0$) did not actually degrade performance. Analysis on individual queries showed that family history did indeed have little effect on every query — it was not the case that family history harmed and improved an equal number of queries, leading to an overall performance comparable to the baseline.

In summary, these results show that negation has overall a detrimental effect on retrieval effectiveness; while, family history does not affect retrieval in a significant way.

RQ3: Does negation always harm (and should we always exclude or negatively weight negated content)?

The overall retrieval results motivate always negatively weighting negated content; the best P@10 settings being $\lambda_{neg} = -0.7$. To verify whether this holds for each of the 81 queries, we analyse the effect of negation weighting on individual queries. Figure 2 shows the change in P@10 for each query when using the best setting of $\lambda_{neg} (-0.7)$ compared to the baseline. In the figure, positive values indicate queries that benefit from negation handling, while negative values indicate queries harmed by negation handling. The results confirm that overall negation handling has positive effects; however, there were a number of queries that were harmed or not affected ($\Delta P@10 = 0$). This finding, i.e., the difference between overall effectiveness and individual query effectiveness, explains the mixed results of previous studies. Studies that considered overall results concluded that negation harms effectiveness [6], while other studies that used a different set of individual queries found that negation had less of an adverse effect [5].

The analysis on individual queries suggests that a single parameter setting across all queries is sub-optimal. We thus perform a full exploration of the parameter space to determine the optimal parameter value on a per-query basis. Figure 3 presents the optimal $\lambda$ value for each of query.

For normal content we observe that most queries are optimised with $\lambda_{nor} = 1.0$, although there are still a number of queries for which the best effectiveness is achieved by reducing the importance of normal content.

For negated content, results are mixed: a number of queries
Figure 3: Optimal $\lambda$ (based on P@10) for each query. Optimal weighting of negated content is polarised (between -1.0 and 1.0), showing that negated content can sometimes aid effectiveness.

<table>
<thead>
<tr>
<th>Method</th>
<th>bpref</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best fixed overall weighting</td>
<td>0.4017</td>
<td>0.4975</td>
</tr>
<tr>
<td>Optimal per-query weighting</td>
<td>0.4526</td>
<td>0.6271</td>
</tr>
<tr>
<td>($+13%$)†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+$26%$)†</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Potential improvements for optimal per-query weighting of different content, compared to the best fixed overall weighting baseline ($\lambda_{nor} = 1.0$, $\lambda_{neg} = -1.0, \lambda_{fh} = 0.8$). † = paired t-test, $p < 0.01$.

are most effective with negative weighting, while only a few queries benefit from negation removal ($\lambda_{neg} = 0.0$), and a large number of queries ($\approx 30\%$) are most effective with positive negation handling ($\lambda_{neg} = 1.0$). The overall results presented in the previous sections highlighted the preference for negative weighting over negation removal. However, the exploration of the optimal parameters illustrated here shows that there is a large number of queries that benefit from including negation via positive weighting. These might be documents that contained both affirmed and negated references to a query term where the overall status was affirmed. For example, a patient who tested positive to a condition at the beginning of their hospital admission, were treated, and a further test showed the condition was no longer present. Additionally, these could be examples where the NegEx algorithm incorrectly annotated a portion of text as negated when it should have been marked as normal.

For family history, the optimal parameter settings show that generally this can be treated like normal content.

The results show that the optimal parameter settings vary significantly between queries. Negation does not always harm performance: in many queries a positive weight should be assigned to negated content. The optimal weights also vary for both normal and family history content.

**RQ4: Can retrieval effectiveness be significantly improved by finding an optimal mix of normal, negated and family history content?**

Given the variability in optimal parameter values, a per-query approach may be beneficial. To quantify the possible advantage of an adaptive strategy that optimally mixes content types, we adjust weighting parameters on a per-query basis using the best settings determined in the previous section. Table 2 shows the retrieval results using this optimal per-query weighting as compared to the best fixed overall weighting baseline ($\lambda_{nor} = 1.0, \lambda_{neg} = -1.0, \lambda_{fh} = 0.8$).

Optimal per-query weighting can significantly improve retrieval effectiveness, more so in the top-ranked results (P@10). This again highlights how negated content does not always harm effectiveness and may be important to include for certain queries. It also motivates further research into an adaptive per-query estimation of the weighing parameters. Methods for estimating per-query parameters are often investigated in IR [7]. A supervised machine learning method may be employed, with the optimal parameter settings identified here supplied as training data. In addition, a set of features must be selected which might indicate which content type is most important given the query; some useful features to choose in this regard might include statistics related to frequencies of occurrences of terms both in normal and negated content, and collection-level statistics, for example, how rare a term and how often does it occur negated.

5. CONCLUSIONS

This study provides an understanding of how and why negation affects clinical IR: overall, negation harms retrieval, family history has little effect. We show that assigning a negative weight to negated content is more effective than the common practise [9, 8] of removing or ignoring this content. However, on an individual query level, negated content can be beneficial and therefore negated content within a document should not be ignored. The difference between overall retrieval effectiveness and individual query effectiveness explains the mixed results of previous studies in this area [5, 6]. Considering negated, family history and normal content separately is flexible and effective for handling these different content types. This approach can easily be applied to other content types beyond negation and family history. An analysis of the optimal weighting showed that significant improvements are possible if the right content mix is chosen. The analysis also revealed possible limitations of our study in that some errors could come from the NegEx and ConText algorithms (while NegEx does have F-measure 90% [2, 4], ConText has not been robustly evaluated). This may explain why no significant improvements were found when tuning the family history weights, although this may also indicate that family history is less important than negation. Future work will investigate an adaptive per-query method to automatically derive the importance of different content type in a clinical document to improve retrieval.

**References**