Generating Clinical Queries from Patient Narratives

A Comparison between Machines and Humans

Bevan Koopman
CSIRO
Brisbane, Australia
bevan.koopman@csiro.au

Liam Cripwell*
CSIRO
Brisbane, Australia
ljcripwell@gmail.com

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

ABSTRACT

This paper investigates how automated query generation methods can be used to derive effective ad-hoc queries from verbose patient narratives. In a clinical setting, automatic query generation provides a means of retrieving information relevant to a clinician, based on a patient record, but without the need for the clinician to manually author a query. Given verbose patient narratives, we evaluated a number of query reduction methods, both generic and domain specific. Comparison was made against human generated queries, both in terms of retrieval effectiveness and characteristics of human queries. Query reduction was an effective means of generating ad-hoc queries from narratives. However, human generated queries were still significantly more effective than automatically generated queries. Further improvements were possible if parameters of the query reduction methods were set on a per-query basis and a means of predicting this was developed. Under ideal conditions, automated methods can exceed humans. Effective human queries were found to contain many novel keywords not found in the narrative. Automated reduction methods may be handicapped in that they only use terms from narrative. Future work, therefore, may be directed toward better understanding effective human queries and automated query rewriting methods that attempt to model the inference of novel terms by exploiting semantic inference processes.

CCS CONCEPTS

•**Information systems** → *Expert search;*

ACM Reference format:
Bevan Koopman, Liam Cripwell, and Guido Zuccon. 2017. Generating Clinical Queries from Patient Narratives. In *Proceedings of SIGIR '17, Shinjuku, Tokyo, Japan, August 07-11, 2017,* 4 pages.
DOI: http://dx.doi.org/10.1145/3077136.3080661

1 INTRODUCTION

An electronic patient record is an invaluable source of information in clinical scenarios. Beyond its immediate use of describing a patient, it provides a reference for retrieving auxiliary information related to that patient such as relevant medical literature or clinical trials for which that patient may be eligible [5]. It is desirable to automatically initiate a search to retrieve such information from a patient record; however, patient records are verbose and using the entire record as an ad-hoc query is not effective [7]. Thus, in this paper, we investigate methods for automatically generating ad-hoc clinical queries from verbose patient reports.

Our main research questions are: 1) Can a verbose patient narrative be reduced to a more effective ad-hoc query? and 2) How does the effectiveness of this query compare with human generated ad-hoc queries?

2 RELATED WORK

The clinical task: Generating clinical queries from verbose patient narratives has previously been attempted as part of the TREC Clinical Decision Support (CDS) track [11]: given a verbose patient narrative, retrieve PubMed articles in a clinical decision support setting. Effective teams typically applied a form of implicit query reduction by weighting terms from the patient narrative [2].

A similar task to TREC CDS involved using the same verbose patient narratives but retrieving clinical trials for which that patient may be eligible [7]. This test collection is of particular relevance as a number of query variations are provided for each topic: the verbose patient narrative, a human provided summary and a number of ad-hoc queries provided by clinicians. These ad-hoc queries therefore provide us with the human benchmark against which any automatically generated query can be evaluated.

Dealing with verbose queries: Previous work has specifically tackled the problem of dealing with verbose queries. Kumaran and Carvalho [7] approached the problem by generating shorter subqueries from the initial query and training a classifier to predict the quality of a given subquery based on various predictive measures [9]. Bendersky and Croft [3] developed a technique for automatically extracting key concepts from verbose queries that had the most impact on retrieval effectiveness. Both these two techniques relied on generating permutations of sub-queries. A key difference between these previous methods and this study was the length of the original verbose queries: Kumaran and Carvalho used topics 3–30 terms in length, Bendersky and Croft used topics 12–49 terms in length, while our patient reports were 39–204 terms in length. For such long queries, some predictive measures were infeasible; for example, generating all possible sub-queries for a 200 term query is intractable (200! combinations). Finally, it is unclear how these general methods translate to the nuances of medical IR [5].

Generating clinical queries: Specific to the medical domain, Soldaini et al. experimented with query reduction techniques for

*Work completed as part of an internship at CSIRO while a student at QUT.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '17, Shinjuku, Tokyo, Japan
© 2017 ACM. 978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080661

searching medical literature [12], including some of those from Kumaran & Carvalho [9].

Koopman et al. [6] experimented with a concept-based information retrieval approach to medical record search using the UMLS medical thesaurus. The experiment showed that queries and documents that are reduced to contain only their medical concepts proved effective. This method relies on effectively identifying medical concepts from free-text; a task that is possible using specialist medical information extraction systems [13].

Understanding effective clinical queries: In order to implement an automated query generation method, it is important to identify the hallmarks of an effective query. This was investigated in Anonymised [1] where the query generation process of humans was examined in detail. The results of that study showed that the most effective queriers were those that inferred novel keywords not present in the original patient narrative. These findings suggest that automated methods which merely reduce the query to contain a subset of its original terms (like those of Kumaran&Carvalho [7] and Bendersky&Croft [3]) may be limited in terms of potential effectiveness. Our empirical investigation answers whether this is the case.

3 METHODS

The first sub-section outlines specific methods for generating shorter ad-hoc queries from a verbose patient narrative. These basic methods are extended to a per-query adaptive approach in the next sub-section. We also outline specific methods for analysing queries to understand more about how automatically generated queries compare with human generated queries.

3.1 Automatic Query Reduction Methods

Proportional Inverse Document Frequency (*IDF-r*): Terms in the original patient narrative were ranked according to inverse document frequency; a proportion of the top ranked IDF terms were retained. This proportion, denoted r, was varied from $1/|D|$ to 1 where $|D|$ was the total number of terms in the patient narratives.[1] The model was run and evaluated with r at all values between 0.01 and 1.0 with increments of 0.01. We denote this *IDF-r*.

Reduce to only UMLS Medical Concepts (*UMLS* & *Tasked-UMLS*): A model was developed that identified and retained only medical related terms from the original patient narrative. Medical terms were identified as those belonging to the UMLS medical thesaurus.[2] Medical terms were identified using QuickUMLS [13] — an information extraction systems that maps free-text to UMLS concepts. We denote this model *UMLS*.

A variant of the UMLS model was also implemented to perform a further reduction to contain only *Diagnosis*, *Treatment* or *Test* related terms. This choice is based on prior studies that show medical professionals typically pose clinical questions around these three types [4] and these form the basis of the queries in TREC CDS [11]. We denote this model *Tasked-UMLS*.

[1] A top-k variant to the *IDF-r* model was also implemented to reduce the topic to include a fixed k terms with highest IDF values; however, the results for this were less reliable than *IDF-r* due to the fact that the lengths of the patient narratives differed considerably.

[2] https://www.nlm.nih.gov/research/umls/

Combined model *UMLS+IDF-r*: Here the original patient narrative was first reduced to contain only medical terms using the *UMLS* model and then a proportion of terms retained using the *IDF-r* model. We denote this model *UMLS+IDF-r*.

3.2 Per-query Reduction via Query Performance Predictors

An important consideration for the aforementioned query reduction methods was how much to reduce the query by (as indicated by query reduction proportion parameter, r). We hypothesised that because topics differed considerably in both length and content, a global setting of r would have been sub-optimal (this was empirically validated in our experiments). Thus, it was desirable to determine the query reduction proportion on a per-query basis. To do this we utilised Query Performance Predictors (QPPs) [9]. Specifically, queries were generated for $r = 0.01..1.0$ in step of 0.01. For each generated query a number of QPPs were calculated. The specific QPPs used were:

Inverse Document Frequency (IDF): This was calculated and averaged across all query terms. $IDF_w = \log \frac{1+N}{n(w)}$ where N is the total number of documents in the collection and $n(w)$ is the collection frequency of term w.

SCQ: A measure of how similar a query was to the collection as a whole; averaged across all query terms. $SCQ_w = (1 + \ln \frac{n(w)}{N}) \times \ln(1 + \frac{N}{N_w})$ where N_w is the document frequency of w.

Inverse Collection Term Frequency (ICTF): This was calculated and averaged across all terms. $ICTF_w = \log_2 \frac{n(w)}{T}$ where T is the total number of terms in the collection.

Query Scope (QS): A measure of the size of the retrieved document set relative to the collection size: $QS = -\log \frac{n_Q}{N}$, n_Q is the number of documents that contain at least one of the query terms.

The correlation between these predictors and the retrieval effectiveness of the queries was examined. In addition, the QPPs were used as features in a model to prediction the value of r; i.e., given a particular topic (patient narrative), determine what query reduction proportion should be applied to it in order to maximise retrieval effectiveness. Training data was obtained by selecting, for each query topic, the best setting of r according to precision @ 5 (P5). This resulted in a total of 1289 topic, query pairs. (Note that for many queries there were many values of r with the same P5; hence the large number of training examples.) The training data was stratified into four folds according to topic id (60 topics divided into folds of 15 topics). A Generalized Linear Model was then trained to predict r based on the QPPs; this was done via 4-fold cross validation. Finally, the predicted values of r were used in *IDF-r* and *UMLS+IDF-r*, and P5, mean reciprocal rank (MRR) and INST calculated.

3.3 Query Understanding Methods

Analysis of how clinician formulate ad-hoc queries from patient narratives has shown that they sometimes selected keywords from the narrative and sometimes inferred novel terms not found in the narrative [1] . Here we consider the overlap of keywords in the clinician's ad-hoc query and corresponding narrative in order to better understand how clinicians formulated their queries.

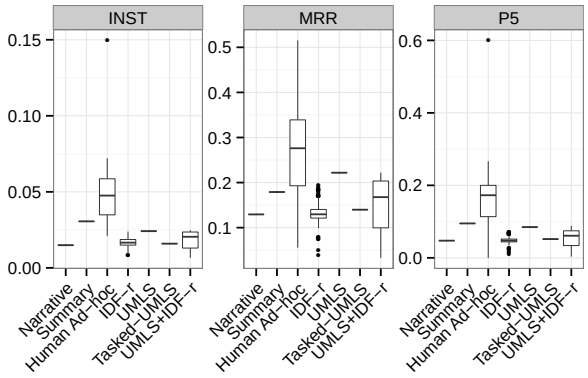

Figure 1: Results for baselines and reduction methods.

The overlap of an ad-hoc query Q is defined as the proportion of keywords in Q that were contained in its narrative text, T: $overlap(T,Q) = \frac{|T \cap Q|}{|Q|}$. The automated query reduction methods outlined in this study clearing were limited to selecting only keywords from the narrative (overlap = 1.0). We, therefore, investigate how much of a handicap this was in comparison to human generated queries containing novel keywords.

3.4 Experimental Setup

Empirical evaluation was performed using a clinical trials test collection [7]. The collection contained 204,855 publicly available clinical trial documents, which we indexed using ElasticSearch with stemming and punctuation removal.[3]

The collection also contained 60 query topics. Each topic contained three different query variations: i) verbose patient case narratives (78 terms per topic); ii) shorter patient case summaries of the patient case narrative (22 terms per topic); and iii) short ad-hoc queries (4.2 terms per topic) provided by clinicians [7]. The narratives represented the original patient narrative (to which query reduction was applied). The shorter summary represented a human benchmark for summarising the narrative. The ad-hoc queries (n=489) represented a human benchmark against which automated methods could be compared.

After query reduction was applied to the narrative, the reduced queries were issued to ElasticSearch and their effectiveness evaluated using P5, MRR and INST (the evaluation measures for this test collection [7]). In addition, the full narrative, summary and ad-hoc queries were also evaluated as comparison baselines/benchmarks. Statistically significant differences in retrieval effectiveness was determined using a paired t-test.

4 RESULTS & DISCUSSION

The retrieval results for the different query reduction methods and comparative baselines and benchmarks are shown in Figure 1. We observe that issuing the entire patient narrative exhibited the poorest retrieval effectiveness. This motivates the develop of specific query reduction methods. The shorter human-generated summaries were more effective than the narrative ($p = 0.030$). This finding highlights that a general reduction of query terms had a positive effect on retrieval. However, the human ad-hoc queries proved far more effective (statistically significant over all other methods). Humans were able to derive specific query keywords that led to more relevant results being retrieved (more on this later). This showed that although a summarisation method had the potential to improve effectiveness, short, ad-hoc keyword queries were still the most effective.

Query reduction via *IDF-r* proved to be effective for specific settings of r. *IDF-r* showed a significant increase in effectiveness in comparison to the narratives ($p = 0.040$) when an appropriate query reduction proportion (r) was chosen. Note that the boxplot shows the effectiveness for all settings of r, many of which would obviously be sub-optimal (e.g., $r = 0.01$ where only 1% of terms were retained). The results for *IDF-r* showed that the removal of less informative terms was a simple but effective means of improving retrieval effectiveness.

Reducing the narrative to contain only medical terms via the *UMLS* method proved effective over searching using the narrative ($p = 0.031$) but not over using the summary ($p = 0.395$). The *UMLS* results showed that simply removing non-medical terms from the narrative was a very good reduction method. *UMLS* seemed to produce better results than the *IDF-r*, although these were not significant ($p = 0.088$). Based on the positive results of the *IDF-r* and *UMLS*, a combined *UMLS+IDF-r* method was evaluated. However, *UMLS+IDF-r* was not statistically significantly different from *UMLS* ($p = 0.568$) or *IDF-r* ($p = 0.072$). However, the *UMLS+IDF-r* had the advantage of having similar effectiveness but with far fewer query terms.

4.1 Understanding human queries

The results from Figure 1 also show that human ad-hoc queries were volatile: they had the greatest variation in effectiveness. While all the best performing queries were ad-hoc, there were also ad-hoc queries that were among the worst performing. Additional analysis is required to determine the characteristics of good vs. bad ad-hoc queries. These findings may help in the development of effective automatic query generation methods. This is left to future work.

Comparing the keywords from ad-hoc queries with those of the patient narrative, it was found that 49% of all queries had an overlap of 0.00; i.e., the ad-hoc query contained no common terms with the narrative. The mean overlap was only 0.26. This indicated that clinicians chose to formulate their own query terms rather than select those from the patient narrative. The inferring of novel query terms, particularly those related to medical treatments, has been found to correlate with higher retrieval effectiveness [1]. Query reduction methods may be handicapped, therefore, by the fact that they source keywords from the narrative alone. We empirically evaluate this in the coming sections and consider further the issue of inferring novel query keywords.

4.2 Query reduction proportion sensitivity

Figure 2 shows the sensitivity to retrieval effectiveness of the query reduction proportion parameter, r ($r = 1.0$ represents the original patient narrative). In general, reduction via IDF proved effective (over the narrative baseline) when the narrative was reduced to approximately 25% of its original size.

[3]ElasticSearch version — 5.2.0 https://www.elastic.co/downloads/elasticsearch.

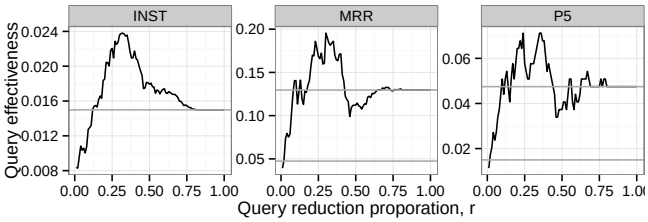

Figure 2: Query effectiveness for different query reduction proportions, r using the *IDF-r*.

Table 1: Retrieval results when predicting query reduction proportion. Percentages show improvements and † show significance when compared to best global r.

	P5	MRR	INST
Human adhoc	0.1521	0.2878	0.0476
Narrative	0.0508	0.1312	0.0149
Summary	0.0949	0.1790	0.0306
Average global r	0.0551	0.1513	0.0184
Best global r	0.0881	0.2220	0.0247
Per-query r — QPP	0.0861 (-2%)	0.2432 (+10%)†	0.0268 (+9%)
Per-query r — Oracle	0.1457 (+65%)†	0.3679 (+66%)†	0.0368 (+49%)†

The results for both *IDF-r* and *UMLS+IDF-r* were all reported with a global reduction proportion, r, across all topics (i.e., across all patient narratives). Since narratives varied greatly in both length and in content, a global reduction proportion may have been quite sub-optimal. Thus, we investigated the effect of predicting r on a per-query basis.

4.3 Predicting query reduction proportion

A Generalised Linear Regression Model was used to predict an appropriate value of r for a given query using the QPP measures of Section 3.2 as features. The results are shown in Table 1. We also report the "oracle" results indicating the retrieval effectiveness if the best value of r was chosen on a per-query basis. Significant improvements in MRR were found when predicting r on a per query basis; no significant differences were found for P5 and INST. This is in contrast to the oracle results that showed considerable improvement if the correct reduction proportion was chosen. Clearly, there is considerable room for improvement. The chief area of focus in this regard is the establishment of a richer set of features for the prediction of per-query r values. Particular points of inquiry would be in the evaluation of medical specific features, such as mentions of particular diseases affecting the patient, permanent demographic information (age, gender) and negated content (e.g., "no fever"). A better understanding of what constituted an effective human query would help to inform such features.

The oracle results showed significant improvement over those produced by maintaining a static, global r value. This highlights that query reduction should ideally be done on a per-query basis. In addition to this, the oracle results were much higher than those of the human generated summaries, showing that automated query generation can improve upon human summarisation. Finally, the oracle results show comparable performance with the human ad-hoc queries. In the case of MRR, the automated query generation methods was statistically significantly better than the human ad-hoc queries ($p = 0.041$). Even though the query reduction method only used terms from the original narrative, the oracle results showed that the right query reduction method was in line with human generated queries that include novel terms. Given that human queries containing novel terms showed greater effectiveness [1], it follows that automated methods for inferring such terms should be investigated. Common query expansion methods are relevant here. However, there are also a number of retrieval techniques, some specific to the medical domain, that attempt to model the inference of novel terms by exploiting semantic inference processes [8, 10, 14].

5 CONCLUSION

Query reduction was an effective means of generating ad-hoc queries from verbose patient narratives. Effective query reduction methods included those that retained only medical terms and a proportion of high ranking IDF terms. Query reduction could be even more effective if the query reduction proportion was determined on a per-query basis. Using standard query performance predictors as features resulted in only minor improvements. However, if an effective query reduction proportion can be found then significant improvements are possible, approaching or exceeding human generated queries.

Human generated queries varied widely in effectiveness. An analysis of human queries showed that many contained novel terms not found in the patient narrative. Queries with novel terms have previously shown to be more effective. Query reduction method may, therefore, be handicapped in that they only source keywords from the patient narrative. Future work, therefore, may be directed toward better understanding effective human queries and in automated retrieval methods that attempt to model the inference of novel terms by exploiting semantic inference processes.

REFERENCES

[1] Anonymised. 2017. Anonymised. *JASIST* To appear (2017).
[2] Saeid Balaneshinkordan, Alexander Kotov, and Railan Xisto. 2015. WSU-IR at TREC 2015 Clinical Decision Support Track: Joint Weighting of Explicit and Latent Medical Query Concepts from Diverse Sources. In *TREC*.
[3] Michael Bendersky and W. Bruce Croft. 2008. Discovering Key Concepts in Verbose Queries. In *SIGIR*. Singapore, 491–498.
[4] J.W. Ely, J.A. Osheroff, P.N. Gorman, M.H. Ebell, M.L. Chambliss, E.A. Pifer, and P.Z. Stavri. 2000. A taxonomy of generic clinical questions: classification study. *BMJ* 321, 7258 (2000), 429–432.
[5] William Hersh. 2008. *Information retrieval: a health and biomedical perspective*. Springer Science & Business Media.
[6] Bevan Koopman, Peter Bruza, Laurianne Sitbon, and Michael Lawley. 2011. AEHRC & QUT at TREC 2011 Medical Track : a concept-based information retrieval approach. In *TREC*. NIST, Gaithersburg, USA, 1–7.
[7] Bevan Koopman and Guido Zuccon. 2016. A Test Collection for Mathcing Patient Trials. In *SIGIR*. Pisa.
[8] Bevan Koopman, Guido Zuccon, Peter Bruza, Laurianne Sitbon, and Michael Lawley. 2015. Information Retrieval as Semantic Inference: A Graph Inference Model applied to Medical Search. *Information Retrieval* 19, 1 (2015), 6–37.
[9] Giridhar Kumaran and Vitor R. Carvalho. 2009. Reducing Long Queries Using Query Quality Predictors. In *SIGIR*. Boston, USA, 564–571.
[10] Nut Limsopatham, Craig Macdonald, and Iadh Ounis. 2013. A Task-Specific Query and Document Representation for Medical Records Search. In *ECIR*. 747–751.
[11] Kirk Roberts, Matthew S Simpson, Ellen Voorhees, and William R Hersh. 2015. Overview of the TREC 2015 Clinical Decision Support Track. In *TREC*.
[12] Luca Soldaini, Arman Cohan, Andrew Yates, Nazli Goharian, and Ophir Frieder. 2015. Retrieving Medical Literature for Clinical Decision Support. In *ECIR*.
[13] Luca Soldaini and Nazli Goharian. 2016. Quickumls: a Fast, Unsupervised Approach for Medical Concept Extraction. In *MedIR Workshop, SIGIR*.

[14] Wei Zhou, Clement Yu, Neil Smalheiser, Vetle Torvik, and Jie Hong. 2007. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *SIGIR*. 655–662.