An Analysis of Query Difficulty for Information Retrieval in the Medical Domain

Lorraine Goeuriot
CNGL, School of Computing
Dublin City University
Ireland
lgoeuriot@computing.dcu.ie

Liadh Kelly
CNGL, School of Computing
Dublin City University
Ireland
lkelly@computing.dcu.ie

Johannes Leveling
CNGL, School of Computing
Dublin City University
Ireland
jleveling@computing.dcu.ie

ABSTRACT

We present a post-hoc analysis of a benchmarking activity for information retrieval (IR) in the medical domain to determine if performance for queries with different levels of complexity can be associated with different IR methods or techniques. Our analysis is based on data and runs for Task 3 of the CLEF 2013 eHealth lab, which provided patient queries and a large medical document collection for patient centred medical information retrieval technique development. We categorise the queries based on their complexity, which is defined as the number of medical concepts they contain. We then show how query complexity affects performance of runs submitted to the lab, and provide suggestions for improving retrieval quality for this complex retrieval task and similar IR evaluation tasks.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval—*Query formulation*; H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing—*Linguistic processing*

Keywords

Medical Information Retrieval; Query Analysis; Evaluation Benchmark

1. INTRODUCTION

Information retrieval (IR) evaluations following the TREC-style tradition typically focus on comparative evaluation of systems and methods, but often put too little emphasis on post-hoc analysis of the task, the associated data, or the submitted runs. In this paper we investigate results of an IR evaluation initiative at CLEF (Cross-Language Evaluation Forum) 2013, the ShARe/CLEF eHealth Evaluation Lab[1] (short CLEF eHealth). Specifically, we analyse Task 3, which is concerned with improving IR systems supporting laypeople in searching for and understanding their health information [2].

[1] http://clefehealth2014.dcu.ie/

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609496.

The specific use case for the evaluation lab is as follows: before leaving the hospital, a patient receives a discharge summary. This describes the diagnosis and the treatment that they received in the hospital. The first task considered in CLEF eHealth aims at extracting names of disorders from the discharge summaries, while the second task requires normalisation and expansion of abbreviations and acronyms present in the discharge summaries. The use case then postulates that, given the discharge summaries and the diagnosed disorders, patients often have questions regarding their health condition. The goal of the third task, a medical IR task, is to provide valuable and relevant documents to patients, so as to satisfy their health-related information needs.

Surprisingly, in this task, no team managed to outperform the strong BM25 baseline provided by the lab [2]. In this paper we examine the topics provided by the organizers for this task, and define levels of query complexity based on the number of concepts in a query. We manually annotate the topics with their complexity category and analyse the performance of participants' runs (i.e. the baseline and best performing run) on these query categories.

The contributions of this paper relate to: 1) analysis of the relationship between query complexity and IR effectiveness; 2) analysis of the performance of different IR techniques through categorisation (based on retrieval technique employed), grouping and analysis of teams baseline runs; 3) analysis of the performance of the best runs across topics with different levels of difficulty; 4) analysis of patterns in the official runs to isolate the impact of individual techniques, methods, or external resources on IR effectiveness.

2. RELATED WORK

The best known analysis of an IR evaluation is the Reliable Information Access (RIA) workshop [3, 9], where retrieval results for different runs and systems were analysed manually to detect weaknesses and system failures in IR systems. One of the main findings is that most systems suffer from the same errors. Harman and Buckley conclude that "it may be more important for research to discover what current techniques should be applied to which topics, rather than to come up with new techniques" [3].

The Robust Track at TREC[2] [11] focused on queries that are difficult for typical systems, aiming to improve the consistency of retrieval technology. This track has resulted in considering evaluation metrics such as the geometric mean average precision for IR when consistent IR effectiveness across all queries is important.

Armstrong et al. [1] have shown that there is very little improvement over strong baselines for publications describing experiments on TREC ad-hoc retrieval. Results for Task 3 of CLEF eHealth show that there is no significant difference in performance metrics

[2] http://trec.nist.gov/data/robust.html

for the "best" submitted run and the baseline experiment [2]. Similarly, results for TRECmed suggest that few systems will outperform a strong baseline [12, 6].

Other related research on improving IR evaluation examined minimizing efforts for relevance assessment by dynamically creating the set of pooled documents [8], determining the quality of test collections [10], or investigating how to automatically predict query performance [5, 4], and exploit this information automatically.

In this paper, we analyse system performance (and IR model performance) for queries with different levels of complexity to determine if particular IR systems, models, methods, or resources can improve a particular subset of topics (i.e. in a query category). Our analysis is based on data from the CLEF 2013 eHealth Task 3 medical IR lab, described next.

3. DATA DESCRIPTION

The CLEF 2013 eHealth Task 3 data comprises a document collection, a set of training and test topics, and relevance assessments. The official data has been released for non-commercial use.

Document Collection. The document collection contains around one million documents, i.e. web pages from medical sites. The documents are predominantly health and medicine websites that have been certified by the Health on the Net (HON) Foundation[3], as well as commonly used health and medicine websites such as DrugBank, Diagnosia, and Trip Answers. The documents are provided in the dataset in their raw HTML format along with their uniform resource locators (URLs).

Topics. The topics (extended queries) were manually created by medical experts, based on information contained in hospital discharge reports. The topic set comprises 5 training topics and 50 test topics.

The queries in the collection aim to model those used by laypeople (i.e. patients, their relatives, or other representatives) to find out more about their condition, after they have examined their hospital discharge summary. The discharge summaries used for the task originate from the anonymized clinical free-text notes of the MIMIC II database, version 2.5[4]. Disorders have been identified within discharge summaries and linked to the matching UMLS (Unified Medical Language System) concepts generated in CLEF 2013 eHealth Task 1 [7].

Registered nurses and clinical documentation researchers developed a set of patient queries using the pairs of discharge summary and a disorder (randomly selected among all disorders identified) in order to generate a set of realistic patient queries.

The generated topics contain a classic TREC-style *title* (text of the query), a *description* (longer description of what the query means), a *narrative* (expected content of the relevant documents; and profile of patient), an additional *discharge-summary* field which links to the associated discharge summary, and a *profile* field, containing information about the patient's profile (such as age, gender, and condition).

Relevance Assessments. Relevance assessment was performed by domain experts and IR experts on documents obtained by pooling the top ten documents from three runs submitted by participants to the CLEF 2013 eHealth Task 3, which resulted in a pool of 6,391 documents. A total of 1,878 documents were assessed as relevant, which is 37.56 per topic on average. Details on the relevance assessment process are described in [2].

Submitted Runs. Participants in this task could submit up to seven different runs, including one baseline experiment (not using any additional or external resources), three experiments not relying on information in the discharge summaries, and three experiments without restrictions. We focus our analysis on their baseline experiment and their top-ranked run.

[3] http://www.healthonnet.org/

[4] http://mimic.physionet.org/

4. QUERY AND RUN ANALYSIS

Although the same process was used to build each topic in the task (described in 3), we observed differences among topics. These differences may be due to the fact that the topics are generated, from a highlighted disorder in a discharge summary, by a human estimating what the information need might be. Therefore, some topics may be directly related to a disease, while others enquire about the relationship between two disorders, or symptoms, for example. We thus categorise queries based on complexity, where complexity corresponds to the number of concepts in a query. We define a concept as a specific medical entity. For example, "diabetes mellitus" is a concept, but "disease" is not.

We established formal guidelines for manual topic annotation with category information (the number of concepts the topic title and description contain) and had three researchers annotate the 55 topics, achieving 75% agreement. Based on the main disagreement, the annotators discussed and reviewed the guidelines. After a second step annotation, they achieved 98% agreement. Specifically, they agreed on the number of concepts for all but one query. The main reason for disagreement was the definition of concept. While it seems rather straightforward to distinguish specific medical entities from general ones, some topics were ambiguous. For example, for the query "White blood cell and bacteria", the annotators could not reach any agreement: two annotators considered "bacteria" to be a concept while the third one did not. This query has been removed from the dataset for the analysis described in this paper. The topic distribution for the 50 test queries is as follows: 22 queries contain one concept (1-concept); 22 queries contain two concepts (2-concept); 5 queries contain three concepts (3-concept); and 1 query is ambiguous. For the 5 training queries we had four 1-concept queries and one 2-concept query.

4.1 Topic results on the task baseline

The task organizers provided a baseline experiment, using the BM25 retrieval model with a standard stop-word list containing the Okapi stop-words (222 stop-words) for stop-word removal [2]. The baseline performs two types of document preprocessing: character normalization (i.e. mapping characters with diacritical marks to the equivalent characters without) and word normalization (e.g. correcting frequent spelling errors). Spelling correction is based on a list of 9533 spelling errors from medical documents [6], which was added to a list of 4192 frequent spelling errors compiled from Wikipedia. During indexing, misspelled words are replaced with their corrections from this list. Table 1 shows the results of the task baseline for each of the 50 test topic categories.

The first column provides the precision at 10 (P@10), which is one of the official CLEF eHealth Task 3 evaluation measures, for each topic category, allowing the analysis of documents returned at top rank, while the second gives the number of relevant documents. P@10 is 0.55 for 1-concept topics, 0.36 for 2-concept topics, and 0.48 for 3-concept topics. However, the performance for 3-concept topics is skewed by 2 topics with P@10 of 0.9 and 1, the remaining 3 topics had P@10 less than 0.4. Thus, as expected, complex multi-concept queries obtain lower performance compared to simpler single concept queries or more precisely, it is more difficult to achieve consistent performance for multi-concept queries (hence the outliers).

Table 1: Results of the task baseline on each topic category

	P@10	# Relevant docs retrieved
1-Concept	0.55	28
2-Concepts	0.36	24
3-Concepts	0.48 (0.375)	137 (19)
average	0.46	36

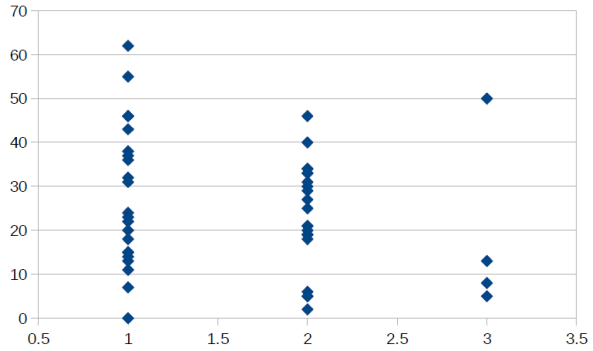

Figure 1: Number of relevant documents per topic (y-axis) for each topic category (x-axis)

Short queries may be expected to obtain better performance as they are less complex; longer queries involving relationships between two concepts might be more difficult to handle for an IR system based on occurrence counting. At the same time, long queries provide much more context, and an IR system is expected to distinguish relevant documents from somewhat relevant ones with such contextual information.

The second column shows the average number of relevant documents per topic. The very high number for 3-concepts is biased by a topic having 610 relevant documents (topic 19), the average number of relevant documents being 19. So this value ranges from 28 documents for 1-concept queries to 19 for 3-concepts ones, which can be explained by the fact that 1-concept queries are typically shorter without much context and are often ambiguous.

Figure 1 shows the distribution of relevant documents per topic category. As can be seen, there does not appear to be any relationship between the volume of relevant documents and topic category, while one could expect 1-concept queries to have many more relevant documents than 2- and 3-concept queries. Therefore the low performances of 2-concept topics cannot be explained by the complexity of the topics resulting in few matching relevant documents.

4.2 Participating teams baselines

Each team participating in the CLEF eHealth task was required to submit a baseline run which did not use any external resources. Figure 2 shows the results of the participating teams baseline runs for each topic category. For comparison, the last group is the CLEF eHealth task baseline described in the previous section. We first observe a pattern in each group. 1-concept topics always perform the best (apart from UTHealth), and 2-concept topics always get the lowest results (apart from Mayo). This is similar to what we observed for the task baseline in 4.1. MEDINFO achieves the highest P@10 on 1-concept topics, Mayo the highest performance on 2-concept topics, and both MEDINFO and MAYO obtain the highest performance on 3-concept topics.

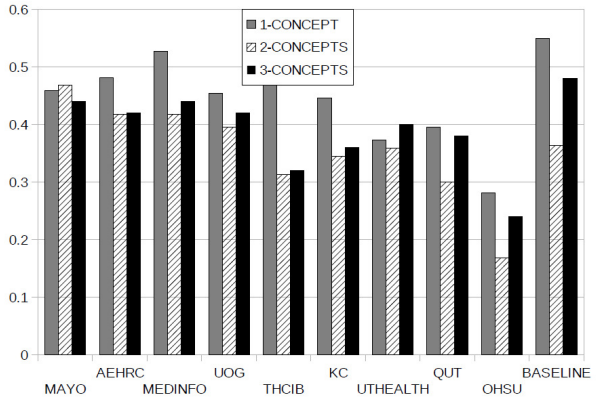

Figure 2: Average P@10 of the participating teams baselines for each topic category

We found that many participating teams used similar IR techniques for their baseline run. Figure 3 shows results of the baselines grouped by IR model. Group 1 uses language modelling (LM) retrieval approaches (teams MAYO, AEHRC, MEDINFO and KC). Group 2 uses the vector space model (VSM) and variants (teams THCIB, UTHEALTH and OHSU). Group 3 uses divergence from randomness (DFR) for retrieval (team UOG). The last team, QUT, used their own proprietary IR model, TOPSIG. As observed for the baseline in the previous section, 1-concept queries obtain the highest P@10 and 2-concept queries the lowest. However, as expected, overall the LM approaches perform better than the weaker VSM baseline. Further, the LM approaches appear to cope much better than the other techniques with 2-concept topics.

We would intuitively expect that longer queries provide more context information as they contain more concepts, compared to possibly ambiguous single concept queries. Similarly, single concept queries may be too generic or unspecific so that they will be associated with a high number of relevant documents, whereas longer queries are more specific and would have less relevant documents.

Results for single concept queries might suffer from missing relevance assessments, following the general observation that the likelihood of finding more (unassessed) relevant documents for queries which already have a high number of relevant documents is high and vice versa. In general, we can observe the same pattern here that for simple queries (1-concept queries), it is easier to obtain a high precision in the top ranks.

We compare groups of runs (grouped together by their IR model) rather than individual teams. We identified the three major IR models used in the submissions for Task 3 as LM, the vector space model (VSM) and its variants, and divergence from randomness (DFR). We excluded retrieval approaches based on proprietary IR models from the analysis described in this paper.

4.3 Participating teams best run

Figure 4 shows the results of the participating teams' best runs for each topic category. The baseline run was the best performing run for three participating teams (team MEDINFO, KC and UTHEALTH). The six other teams obtained an improvement over their baseline using various methods. Two teams obtained a significant improvement in their performance on 1-concept queries (teams MAYO and AEHRC). On 2-concept queries, 4 teams obtained a significant improvement (teams MAYO, AEHRC, THCIB and OHSU). As for the 3-concept queries, two teams obtained bet-

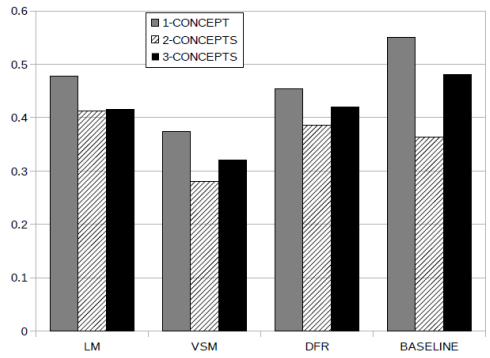

Figure 3: Average P@10 of the participating teams baselines grouped by IR technique, for each topic category

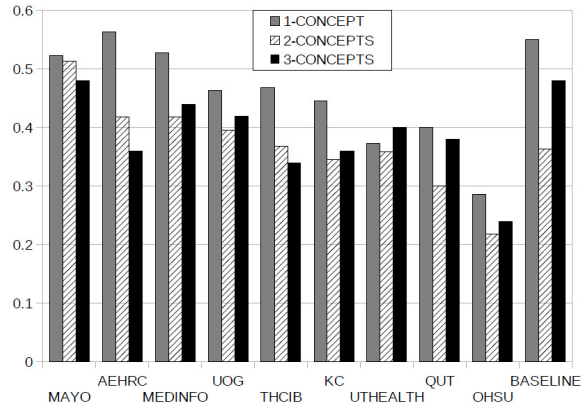

Figure 4: Average P@10 of the participating teams best runs on each topic category

ter results (teams MAYO and THCIB), but one teams performance decreased (team AEHRC).

The first three teams which obtained better performance than their baselines (MAYO, AEHRC and UOG) used very varied approaches.

Team Mayo adds two ranking systems combined to their baseline to obtain their best run. The first is a linear combination of Markov Random Field (MRF) model and a Mixture of Relevance Models (MRM). The second is based on a UMLS CUI-representation of the documents, topics, and discharge summaries. They obtained an improvement for each topic category. It cannot be determined from their runs which part of this system is responsible for the improvement in performance, it is very likely the combination.

Team AEHRC's best run adds topic acronym expansion and spelling correction to their baseline. This greatly improved their retrieval performance on 1-concept topics, but 2-concept topics results are similar to their baseline, and 3-concepts are lower.

Team UOG's best run adds pseudo-relevance feedback, using the DFR Bo1 model, to their baseline run. This addition to the baseline slightly improves retrieval performance for 1-concept queries, and yields similar performance for 2- and 3-concept queries.

5. CONCLUSIONS AND FUTURE WORK

We analysed runs submitted to the CLEF 2013 eHealth evaluation initiative to identify the impact of query complexity on IR performance in the medical domain. Overall, and unsurprisingly, retrieval performance is affected by query complexity. Use of rich retrieval approaches lessens this effect. Best retrieval performance is obtained for simple 1-concept queries. We found that query expansion techniques, such as acronym expansion, while improving 1-concept query retrieval performance, have little effect on multi-concept queries. However, use of sophisticated LM language techniques, as opposed to simpler techniques, decreases the difference in retrieval performance between 1- and multi- concept queries. Use of ontology-based (CUI-based) methods appear to further decrease this difference, as evidenced by the MAYO teams performance on the different categories of queries. For this team however, it is a combination of techniques, as opposed to one single approach that is proving beneficial. Further experiments would be required to tease out the effects of the component parts of their retrieval approach.

As a side note, we observed that teams which used the same baseline IR approach did not obtain the same results. This task (and other tasks) could benefit from providing stricter guidelines on description of baseline experiments, including all parameter settings and all preprocessing steps.

6. ACKNOWLEDGMENTS

This work is supported in part by the Khresmoi project (257528) and SFI (07/CE/I1142) as part of the Centre for Next Generation Localisation at Dublin City University.

7. REFERENCES

[1] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *CIKM 2009*, pages 601–610. ACM, 2009.

[2] L. Goeuriot, G. J. F. Jones, L. Kelly, J. Leveling, A. Hanbury, H. Müller, S. Salanterä, H. Suominen, and G. Zuccon. ShARe/CLEF eHealth Evaluation Lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In *CLEF online working notes*. 2013.

[3] D. Harman and C. Buckley. The NRRC reliable information access (RIA) workshop. In *SIGIR 2004*, pages 528–529. ACM, 2004.

[4] C. Hauff, F. de Jong, D. Kelly, and L. Azzopardi. Query quality: User ratings and system predictions. In *SIGIR '10*, pages 743–744, New York, NY, USA, 2010. ACM.

[5] B. He and I. Ounis. Query performance prediction. *Inf. Syst.*, 31(7):585–594, Nov. 2006.

[6] J. Leveling, L. Goeuriot, L. Kelly, and G. J. F. Jones. DCU@TRECMed 2012: Using ad-hoc baselines for domain-specific retrieval. In *TREC 2012*. NIST, 2012.

[7] S. Pradhan, N. Elhadad, B. South, D. Martinez, L. Christensen, H.Suominen, W. Chapman, and G. Savova. Task 1: ShARe/CLEF eHealth. In *CLEF online working notes*, 2013.

[8] T. Sakai and T. Mitamura. Boiling down information retrieval test collections. In *RIAO 2010*, pages 49–56. CID, 2010.

[9] I. Soboroff. A guide to the ria workshop data archive. *Information Retrieval*, 12(6):642–651, Dec. 2009.

[10] J. Urbano, M. Marrero, and D. Martín. On the measurement of test collection reliability. In *SIGIR '13*, pages 393–402. ACM, 2013.

[11] E. M. Voorhees. The TREC robust retrieval track. *SIGIR Forum*, 39(1):11–20, June 2005.

[12] E. M. Voorhees and W. Hersh. Overview of the TREC 2012 medical records track. In *TREC 2012*. NIST, 2012.