A Study on Manual Query Reformulation for Systematic Medical Reviews
(Extended Abstract)

Giorgio Maria Di Nunzio[1] and Federica Vezzani[2]

[1] Dept. of Information Engineering – University of Padua
[2] Dept. of Linguistic and Literary Studies – University of Padua
giorgiomaria.dinunzio@unipd.it, federica.vezzani@phd.unipd.it

Abstract. Technology-Assisted Review (TAR) approaches are essential to minimize the effort of the user during the search and collect all relevant documents. In this paper, we present a failure analysis based on terminological and linguistic aspects of a TAR system for systematic medical reviews. In particular, we analyze the results of the worst performing topics of the best experiments of the CLEF 2017 eHealth task on Technologically Assisted Reviews in Empirical Medicine. This is an extended abstract of the work presented in [2, 4].

1 Introduction

The large and growing number of published medical studies makes the task of identifying relevant documents for the realization of systematic medical reviews complex and time consuming [6]. As a matter of fact, "it is unlikely that [healthcare providers, researchers, and policy makers] will have the time, skills and resources to find, appraise and interpret all this evidence and to incorporate it into healthcare decisions."[3] For this reason, semi-automatic TAR approaches are essential to minimize the effort of the user during the search and collect all relevant documents [1].

In this paper, we present the ongoing work about a methodology based on linguistic and terminological aspects which, in conjunction with a semi-automatic TAR system, allows us to obtain high recall results from queries written by users without medical skills. This task is time consuming and it requires specific linguistic skills in order to find the best lexical representation of an information need. We intentionally performed our experiment by asking for the support of users who are not experts in the medical field but are familiar with the linguistic domain. The presented methodology is based on a linguistic and terminological approach and it follows a series of well-structured steps used to build effective continuous active learning system for medical domains [3]. Starting from the information need provided by a physician, the approach is based on the following steps: 1) Recognition of technical terms; 2) Extraction of technical terms; 3) Formulation of terminological records; 4) Query rewriting.

IIR 2018, May 28-30, 2018, Rome, Italy.

[3] http://handbook-5-1.cochrane.org

Variant	Query
information need	First rank symptoms for schizophrenia
expert keywords	diagnosis, diagnostic, first rank symptoms, symptom, schizophrenia, FRS, international pilot study, IPSS, schneider, schneiderian, schizophrenics, non-schneiderian
expert readable	Diagnostic accuracy of one or multiple FRS for diagnosing schizophrenia as a psychotic disorder
group keywords	Schizophrenia, schizophrenic, first rank symptoms, Schneider, schneiderian symptoms, FRS, diagnostic criteria, psychopatological, DSM, ICD, pathognomonic, disturb
group readable	Diagnostic specificity of hallucination, delusions, thought interference and other schneiderian symptoms (FRS) for the diagnosis of schizophrenia
individual variant1	Specificity and relevance of the Schneiderian symptoms (FRS) in order to diagnose the schizophrenic psychosis
individual variant2	schneiderian symptoms, FRS, schizophrenia, schizophrenic, diagnostic, psychopathology, pathognomonic, specificity, disturb, ICD, meta analysis.

Table 1: Query reformulation for the topic CD010653

2 Experiments

We present the ongoing experiments of this methodology applied to the CLEF 2017 eHealth Task on TAR in Empirical Medicine[4]. This task focuses on the problem of systematic medical reviews by providing information needs prepared by physicians and the relative relevance judgments. The objective of the task is to retrieve "all" the relevant documents with the least effort. The dataset provided by the task is based on 50 systematic reviews, or topics, conducted by Cochrane experts on Diagnostic Test Accuracy. The dataset consists of: a set of 50 topics (20 training and 30 test) and, for each topic, the set of PubMed Document Identifiers (PIDs) returned by running the query in Pubmed as well as the relevance judgments for both abstracts and documents [5].

We performed two experiments: the first experiment involved two experts in linguistics, while the second experiment involved 90 undergraduate students of a Master Degree in Foreign Languages and Literary Studies. During the first experiment, we tested the feasibility of the query rewriting approach with the collaboration of the two experts in linguistics. Each expert had a different goal: one expert was instructed to describe the original information need with the first type of query variant, i.e. a list of keywords, while the other expert was in charge of the second type of query rewriting, the reformulation of the information need with a humanly comprehensible query (for example, see the first two rows of Table 1). During the second experiment, we divided the 90 student volunteers of the course into 30 groups of 3 people each. Each group was entrusted with a specific information need for the medical field and a goal: to reformulate the initial query, as a group and invididually (see the last four rows of Table 1, respectively), by evaluating specific linguistic aspects in order to give two reformulations according to the above mentioned methodology. The dataset of query reformulation is openly available in a GitHub repository.[5]

[4] `https://sites.google.com/site/clefehealth2017/task-2`

[5] GitHub link to be updated

3 Results

The system used in our experiments implements the AutoTAR Continuous Active Learning (CAL) method proposed by [1]. The system is based on a BM25 weighting scheme which is updated whenever the system identifies a relevant document [2]. The retrieval system has two parameters that can be set to adjust the amount of documents that a physician is willing to review: the *percentage* p of documents over the number of documents retrieved by the original boolean query, the *threshold* t of the number of documents to read. The parameter p is used to find the initial estimates of the probabilities of each term in the ranking phase while t sets the maximum number of documents that a physician is willing to read before the final round of classification. Following the indications given [2], we vary the parameter p from 10 % to 50% and set t equal to 500 and 1000, respectively. For each combination of values of p and t, 10 in total, we produce three types of runs: a run named 'expert' with the query variants produced by the two experts in linguistics, a run named 'group' with with the query variants created by each group of students, a run named 'individual' with the variants written by each student of each group. For the evaluation of our experiments, we used the official scripts provided by the organizers of the CLEF eHealth task[6]. This repository also contains the official results of all the participants to the task, we use these results as a baseline for our analyses. We present the results of the experiments in three parts: a comparison with the official runs of the CLEF 2017 task, an analysis among the top performing runs, a brief failure analysis.

Comparison with CLEF 2017 runs As described by [2], all our runs dominate the Pareto frontier (the best performances in terms of effort vs recall) across all the range of documents shown. In particular, the best runs with threshold $t = 500$ achieve the same recall of the best CLEF run with the same recall using around 20,000 documents less (40,000 vs 60,000), while the best runs with $t = 1000$ achieve almost the same perfect recall of the CLEF run (0.993 vs 0.998) using 25,000 documents less (63,000 vs 88,000).

Comparison Across Runs We performed a Wilcoxon paired signed test for every pair of types of runs (expert, group, individual). The result confirms that there is no statistically significant difference among the performances of the runs. This means that the system performs well even when the query is written by a non expert in the field of medicine.

Low Recall Topics We perform a failure analysis on those topics for which the system did not achieve a recall of 100%. For $t = 500$ and $p = 50\%$ there are only 10 topics that do not achieve a perfect recall. Among these topics, we focus on topic CD010653 since it is the one with the largest difference in performance

[6] `https://github.com/leifos/tar`

among the runs. From a linguistic point of view it is interesting to note the differences between the expert keywords reformulation and the individual variant 2, see Table 1. On one hand, the first reformulation uses a lexical morphological approach: more variants (inflected forms) of the same term are proposed such as *diagnosis, diagnostic* or *schneider, schneiderian, and non-schneiderian*. The individual variant 2, on the other hand, aims at covering the involved semantic sphere: the participant uses terms such as *psychopathology, pathognomonic, specificity, ICD and meta analysis* that are not present in other reformulations. The reformulation approach adopted, the morphological or the semantic one, may therefore have influenced the results of the performance, but we shall analyze in more detail this particular emerging feature in future works.

4 Conclusions

In this paper, we have presented a methodology based on linguistic and terminological aspects which is functional to the query rewriting task. We have applied this methodology to the TARs in Empirical Medicine CLEF 2017 task, so that users without medical skills were able to reformulate 30 specific medical information needs provided by physicians. As future work, we will investigate the semantic and morphological behavior of 5 topics which are part of the dataset for CLEF 2017. Through a more in-depth analysis, we have found that these topics have a relevant latent terminology. In particular, the unexctracted terms are acronyms and inflected forms of nouns and verbs. We therefore propose to focus on the operations of "expansion" and/or "implosion" of acronyms and on the process of lemmatisation of the inflected forms of the terms through their reduction to pure lemmas.

References

1. G. V. Cormack and M. R. Grossman. Scalability of continuous active learning for reliable high-recall text classification. In *Proc. of CIKM'16*, pages 1039–1048, New York, NY, USA, 2016. ACM.
2. G. M. Di Nunzio. A study of an automatic stopping strategy for technologically assisted medical reviews. In *Proc. of ECIR 2018*, pages 672–677. Springer, 2018.
3. G. M. Di Nunzio, F. Beghini, F. Vezzani, and G. Henrot. An interactive two-dimensional approach to query aspects rewriting in systematic reviews. IMS unipd at CLEF ehealth task 2. In *Working Notes of CLEF 2017*. CEUR-WS.org, 2017.
4. G. M. Di Nunzio and F. Vezzani. Using R markdown for replicable experiments in evidence based medicine. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, LNCS, page In Press. Springer, 2018.
5. E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker, editors. *CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview. In Working Notes of CLEF 2017*. CEUR-WS.org, 2017.
6. A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou. Using Text Mining for Study Identification in Systematic Reviews: A Systematic Review of Current Approaches. *Syst. Rev.*, 4(1):5, 2015.