# Using R Markdown for Replicable Experiments in Evidence Based Medicine

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.** In this paper, we propose a methodology based on the R Markdown framework for replicating an experiment of query rewriting in the context of medical eHealth. We present a study on how to re-propose the same task of systematic medical reviews with the same conditions and methodologies to a larger group of participants. The task is the CLEF eHealth Task Technologically Assisted Reviews in Empirical Medicine which consists in finding all the most relevant medical documents, given an information need, with the least effort. We study how lay people, students of a master degree in languages in this case, can help the retrieval system in finding more relevant documents by means of a query rewriting approach.

## 1 Introduction

Systematic medical reviews are a method to collect the findings from multiple studies in a reliable way and are used to inform policy and practice [19]. During the 'screening' of documents, physicians look manually through collections of medical databases in order to identify most (if not all) the relevant documents pertaining the object of the search. In this context, Technology-Assisted Review (TAR) systems help the user to find as much relevant information as possible with reasonable effort [5]. The most successful TAR systems tackle the problem by training a classifier by means of a continuous active learning approach (each time a user reads a new document and judges it relevant or not, this information is immediately given as feedback to the system) [23, 19]. There is also the problem related to how the user form the query in order to restrict the set of documents to be considered. The principal systems in current use are document databases supporting Boolean querying. Reviewers use such systems to incrementally build complex queries that may involve hundreds of terms, with the aim of including the great majority of relevant documents in the answer set. For example, in [14], the authors investigate a hybrid approach, where a Boolean search strategy is used to fetch an initial pool of candidate documents, and ranking is then applied to order the result set.

In this paper, we follow a similar approach to [14] and we present a methodology for replicating an experiment of query rewriting in the context of medical eHealth. In particular, the experiment consists in re-proposing a task previously

performed by a team of researchers for the CLEF eHealth Task 2: "Technologically Assisted Reviews in Empirical Medicine" [7]. The task consists in retrieving all the relevant documents for medical specific domains as early as possible and with the least effort. The main goal of the experiment presented in this paper is to re-propose:

- the same task with
- the same conditions and
- the same methodologies
- to a larger group of participants.

In particular, we want to accomplish the goal of the 'Replication track' of this Task in order to disseminate solid and reproducible results, as also shown by [8].

### 1.1 Replicability Issues in IR Experiments

Replicable and reproducible methods are fundamental research tools because the lack of reproducibility in science causes significant issues for science itself. Research areas in Computer Science using linguistic resources in an extensive way have been addressed this problem in the last years. For example, the most important conferences in Information Retrieval (IR) support this kind of activities [9]: the open source information retrieval reproducibility challenge at SIGIR[3], the Reproducibility track at ECIR since 2016 [10], as well as some Labs at the Cross-Language Evaluation Forum (CLEF) that explicitly have a task on reproducibility, such as CLEF eHealth.[4] In 2018 the three major conferences in IR evaluation, TREC, CLEF and NTCIR made a joint effort to support replicable research through the CENTRE initiative.[5]

The Natural Language Processing (NLP) community has witnessed the same issue. In 2016, the "Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language" at the Language Resources and Evaluation Conference (LREC) encouraged the discussion and the advancement on the reproducibility of research results and the citation of resources, and its impact on research integrity in the research area of language processing tools and resources [2]. In the very recent past, even the second edition of the 4REAL workshop at LREC 2018 aimed to contribute to the advancement of reproduction and replication of research results which are at the heart of the validation of scientific knowledge and scientific endeavor.

> "Everyone agrees that there's a problem: very often, results and conclusions in experimental science and some areas of engineering turn out to be unreliable or false. And everyone agrees that the solution is to put more effort into verifying such results and conclusions, by having other people re-do aspects of the research and analysis [17]".

[3] `https://goo.gl/CePVzY`
[4] `https://goo.gl/WgkqnZ`
[5] `http://www.centre-eval.org`

| **Technical term: RADICULOPATHY** | | | | **Traducente: RADICOLOPATIA** | | | |
|---|---|---|---|---|---|---|---|
| **Formal features** | | | | **Caratteristiche formali** | | | |
| Genre | Noun | | | Genere | s. f. | | |
| Spelling | "th" pronounced as a voiceless dental non-sibilant fricative | | | Ortografia | niente da segnalare | | |
| Tonic accent | /ˌradɪkjʊ'lɒpəθi/ | | | Accento tonico | radicolopatia | | |
| Derivation/Composition | Latin radicula + English -o- + -pathy<br>First Known Use: 1942 | | | Derivazione/Composizione | [comp. del lat. radicŭla (dim. di radix -icis «radice») e di -patia] | | |
| **Language Definition** | **Field** | **Register** | **Sources** | **Definizione in lingua** | **Dominio** | **Registro d'uso** | **Fonti** |
| Dysfunction of one or more spinal nerve roots, characterized by pain and sensory and motor disturbances and often caused by compression | Pathology | Specialized | 1 | Qualsiasi alterazione a carico di una radice nervosa, determinata da cause varie: infiammatorie, compressive, tossiche, malformative, vascolari, ecc. | Patologia | Specializzato | 1 |
| /dysfunction//spinal nerve//root//body//living being/ | | | | /alterazione//radice//nervo//corpo//essere umano/ | | | |
| **Specialized contexts** | **Field** | **Register** | **Sources** | **Proposte di traduzione** | | | |
| This is why neck problems that affect a cervical nerve root can cause pain and other symptoms through the arms and hands (**radiculopathy**), and low back problems that affect a lumbar nerve root can radiate through the leg and into the foot (radiculopathy, or sciatica), thus prompting leg pain and/or foot pain. | Pathology | Specialized | 2 | This is a semantically univocal term: the result is the perfect correspondence of units of sense. For this reason, the italian translating candidate for "radiculopathy" is "radicolopatia". | | | |
| Risk factors for **radiculopathy** are activities that place an excessive or repetitive load on the spine. | Pathology | Specialized | 3 | | | | |
| **Ontological - encyclopedic illustrations** | https://www.google.it/search?q=radiculopathy&rlz=1C1AVNE_enIT611IT612&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiK-tKrwdnYAhVMCewKHUMZD00Q_AUICigB&biw=1093&bih=670#imgrc=xQ8-5mJoQYTZBM: | | | **Illustrazioni ontologiche - enciclopediche** | https://neurourologia.files.wordpress.com/2007/10/radicolopatia-s1.jpg | | |
| **Collocations - phraseology** | RADICULOPATHY+VERB : R. occurs when [...], R. results when [...] | | | **Collocazioni - fraseologia italiana** | Soffrire di radicolopatia | | |
| **References** | | | | **Riferimenti** | | | |
| https://en.oxforddictionaries.com/definition/radiculopathy | | | 1 | http://www.treccani.it/vocabolario/radicolopatia/ | | | 1 |
| https://www.spine-health.com/conditions/spine-anatomy/radiculopathy-radiculitis-and-radicular-pain | | | 2 | https://it.wikipedia.org/wiki/Radicolopatia | | | 2 |
| https://www.medicinenet.com/radiculopathy/article.htm | | | 3 | | | | |

Fig. 1: Multilingual Terminological Record: Radiculopathy

The benefit of reproducibility is evident in cases where faithfully recreating the research conditions is impossible as the variables contributing to a particular instance of field observation are too hard to control in some cases [1]. Clearly, linguists cannot expect their colleagues to replicate data collection conditions, and doing so would not necessarily lead to replicated utterances, but reproducibility is a more realistic goal.

In this paper, in order to describe in detail the process of data preparation, we use the 'literate programming' approach proposed by Donald Knuth [15]. Literate programming helps peers understand and replicate your results, find errors and suggest enhancements and, ultimately, produce better-quality programs.

We used the R Markdown framework[6] since it is considered one of the possible solutions to document the results of an experiment and, at the same time, reproduce each step of the experiment itself. Following the indications given by [11] and the suggestions discussed by [3], we developed the experimental framework in R and we share the source code on Github in order to allow other participants to reproduce and check our results.[7]

## 2 Linguistic Methodology for Query Rewriting

In this section, we outline the linguistic methodology that the participants of the experiment used to for rewrite the original query of the expert in order to capture different senses of the information need and retrieve more relevant documents. We proceeded by the reformulation of an initial query given by an expert by planning our working methodology on the analysis of some linguistic and terminological aspects functional to the process of query rewriting. This approach

[6] `http://rmarkdown.rstudio.com`

[7] `https://github.com/gmdn/CLEF2018`

has contributed to an effective and efficient reformulation for the retrieval of the most relevant documents for the research. The approach is divided into the following steps:

1. Identification of technical terms;
2. Manual extraction of technical terms;
3. Linguistic and semantic analysis;
4. Formulation of terminological records;
5. Query rewriting.

The basis of our methodology for query rewriting is a terminological and linguistic analysis of the initial query formulated by the expert. Given the short information need, we started with the identification of the technical terms, as all the terms that are strictly related to the conceptual and practical factors of a given discipline or activity [16]. Medical language has actually a specialised vocabulary composed of strictly specialised terms referring to this particular domain.

We then proceed with the manual extraction of such technical terms and started to conduct a linguistic and terminological analysis through the implementation of the core of our methodology for query rewriting, that is a new model of terminological record. Terminological records are commonly used in terminology and linguistics as a tool for the collection of terminological and linguistic data referring to a specific concept [12]. The term records proposed to the participants of the experiment is based on the model implemented in a linguistic resource aiming to provide a support eHealth tool for the study of the complexity of medical language from the semantic viewpoint: TriMED [24]. The new model of term record provide information both from a purely linguistic and from a translation point of view. TriMED is actually designed for technical-scientific translators who have the difficult task to decode and then transcode medical information from a source language into a target language. For this reason, the terminological record offers the same kind of information for the technical term and its equivalent in the target language. Figure 1 shows an example of a terminological record for the technical term *Radiculopathy* and its equivalent in Italian *Radicolopatia*.

Focusing on the linguistic aspects, these records provide a broad spectrum of information of the term analysed. We firstly decide to provide all the formal features related to the term that are necessary for its lexical framing:

- Genre;
- Spelling:
- Tonic accent;
- Derivation and composition.

In order to grasp the content of concepts, we provide the definition of the terms through the analysis of the meaning conventionally attributed to them by a community of people sharing the same knowledge and having a common goal. Definitions constitute a structured system of knowledge [18] in order to understand the meaning of a term. We extracted the definitions from reliable resources

as Merriam-Webster Medical Dictionary[8] and MediLexicon[9] in particular for acronyms and abbreviations. Furthermore, we focus on the semantic viewpoint by providing the semic analysis of the term. Semic analysis is a methodology of study used in compositional semantics aiming to decompose the meaning of technical terms (that is the lexematic or morphological unit) into minimum unit of meaning that cannot be further segmented, known as semantic traits or semantic components. The union of multiple semantic traits makes up the meaning of a lexeme [21].

Moreover, participants were required to provide the context of use of the term. This is because, in a such specific domain, the context attributes the semantic value to the term. Participants considered phraseology (collocations in particular) in order to analyse the semantic behaviour of the terms related to their neighbours. Phrasemes are intended as the combinations whose overall meaning does not result from the sum of the meanings of the individual components [6]. Finally, terminological records offer ontological illustrations of the term and some references in order to track the retrieval of information.

With this kind of analysis participants were able to

1. create the basis of knowledge for the domain and the context of study;
2. propose the query variant through two different approaches.

The first variant of the query was a list of keywords that the participants obtained from the semic analysis of the technical terms contained in the initial query. The second variant is instead a human readable reformulation, therefore grammatically correct, and containing the fewest possible number of terms equal to the starting query. This reformulation is therefore made up of synonymic variants, acronyms, abbreviations or periphrases. Participants could exploit the information given by a document that could be relevant or not according to the initial query, the list of term frequencies, document frequency and the boolean query generated by PubMed.[10]

## 3 Experiments

The participants of this experiment were the students of the Master's Degree course in Modern Languages for International Communication and Cooperation of the University of Padua. The 90 students, all of them with different background, were divided into 30 groups of 3 people each. Each group has been entrusted with a specific information need for the medical field. The aim of the students was to reformulate the initial query by evaluating specific linguistic aspects in order to give two reformulations according to the above mentioned methodology. The result is a number of 60 reformulations, that is two variants of queries formulated by each group of students. Hereinafter an example of the two variants proposed by a group of students for a specific information need:

[8] `https://www.merriam-webster.com/medical`
[9] `http://www.medilexicon.com`
[10] `https://www.ncbi.nlm.nih.gov/pubmed/`

– Initial query: *Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain*;
– First variant: *Sensitivity, specificity, test, tests, diagnosis, examination, physical, straight leg raising, slump, radicular, radiculopathy, pain, inflammation, compression, compress, spinal nerve, spine, cervical, root, roots, sciatica, vertebrae, lumbago, LBP, lumbar, low, back, sacral, disc, discs, disk, disks, herniation, hernia, herniated, intervertebral*;
– Second variant: *Sensitivity and specificity of physical tests for the diagnosis of nerve irritation caused by damage to the discs between the vertebrae in patients presenting LBP (lumbago).*

At a later time, we asked to the students to reformulate an individual query different from the two variants previously proposed:

– Individual reformulation: *Patients with pain in the lower back need a check-up for the compression or inflammation of a spinal nerve caused by rupture of fibrocartilagenous material that surrounds the intervertebral disk.*

The first reformulation is therefore a list of keywords that tends to cover as much as possible the semantic sphere affected by the term analysed. The second human-readable reformulation is more focused on providing synonymic variants or acronyms in order to use the least possible number of terms of the initial query. Whereas, the individual reformulation does not follow a precise approach other than that of human interpretation resulting from the approximate study of the subject contained in the query. At the end of the experiment, 28 groups completed the task. We therefore received a total of 28 list of keywords, 28 human-readable reformulation and 66 individual reformulations.

### 3.1 Dataset and System Settings

The dataset provided by the TARs in Empirical Medicine Task at CLEF 2017 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 [13].

The system that retrieves the documents that the user (the physician in our case) has to judge implements the AutoTAR Continuous Active Learning (CAL) method proposed by [4]. The system is based on a BM25 weighting scheme [22] which is updated whenever the system identifies a document for assessment and the relevance judgment (provided with the CLEF dataset) is used as a feedback [20]. The system has only 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.

In our experiments, we used only the relevance judgments of the abstracts and we did not use any training topic to optimize the system. We used the source code provided by [20] for the Continuous Active Learning method [5] to simulate the interaction with a physician who gives a relevance feedback for each abstract retrieved. Following the indications given by the authors, 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 the query variants created by each group of students, a run named 'individual' with the variants written by each student of each group.

## 4 Results

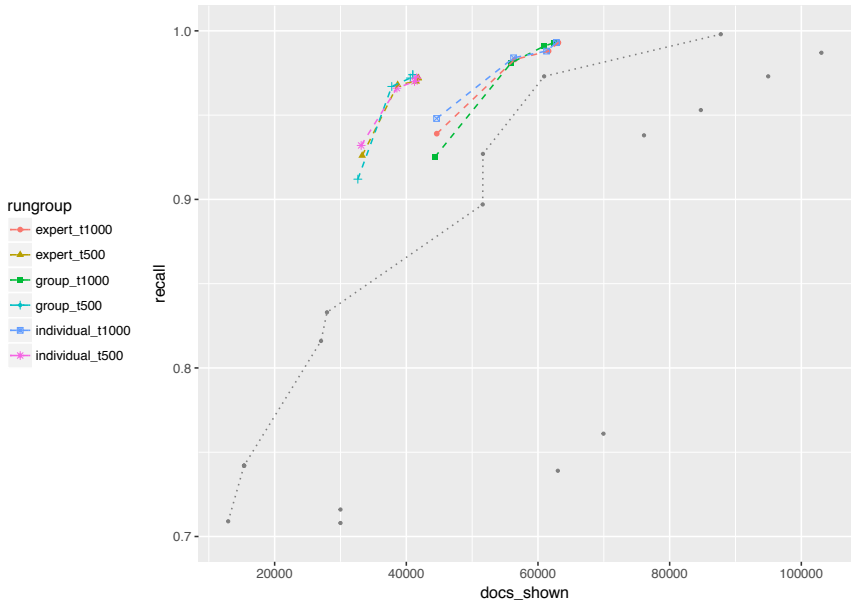
For the evaluation of our experiments, we used the official scripts provided by the organizers of the CLEF eHEalth task.[11] This repository also contains the official results of all the participants to the task, we use these results as a baseline for our analyses. We present the results of the experiments 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** In Figure 2a, we show a comparison between the performances of the runs with threshold $t = 500$ and $t = 1000$ and those of the official runs of CLEF 2017. On the abscissa we have the number of documents shown (the documents that are actually shown to the physician for relevance judgment), on the ordinate the average recall over the 30 topics. The grey points represent the performance of the CLEF 2017 runs, and the dotted grey line the Pareto frontier[12] of the best runs. The coloured lines represent the performance of the three types of runs (expert, group, individual). Each line connects five points relative to the five values of $p$ (from $p = 10$ to $p = 50$). All our runs dominate the Pareto frontier 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).

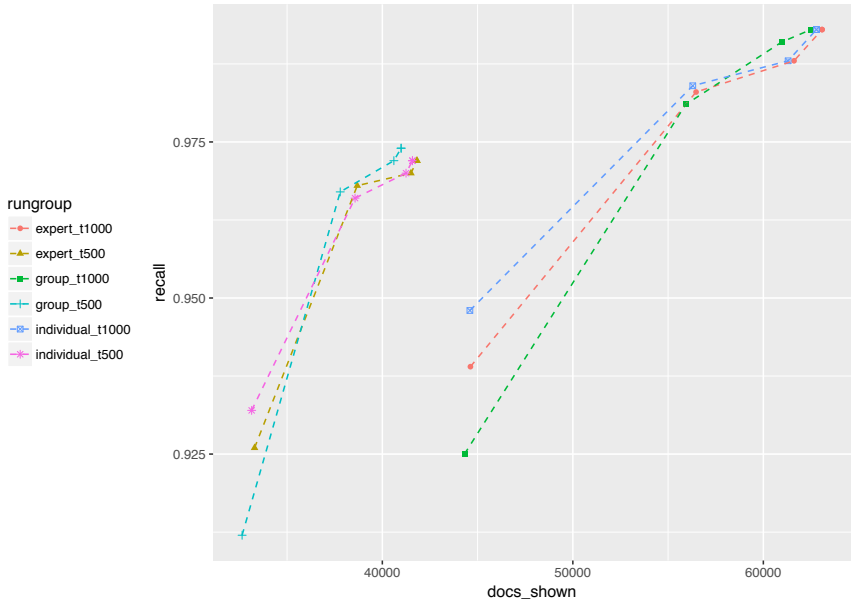
**Comparison Across Runs** In Figure2b, we show a close-up of Figure 2a for the six runs. By increasing $p$ the average recall increases consistently, especially from $p = 10\%$ to $p = 20\%$ and from $p = 20\%$ to $p = 30\%$. When $p = 50\%$

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

[12] `https://en.wikipedia.org/wiki/Pareto_efficiency`




(a) Grey dots are the official CLEF 2017 runs




(b) Close-up of experiments

Fig. 2: Average Recall at total number of documents shown.

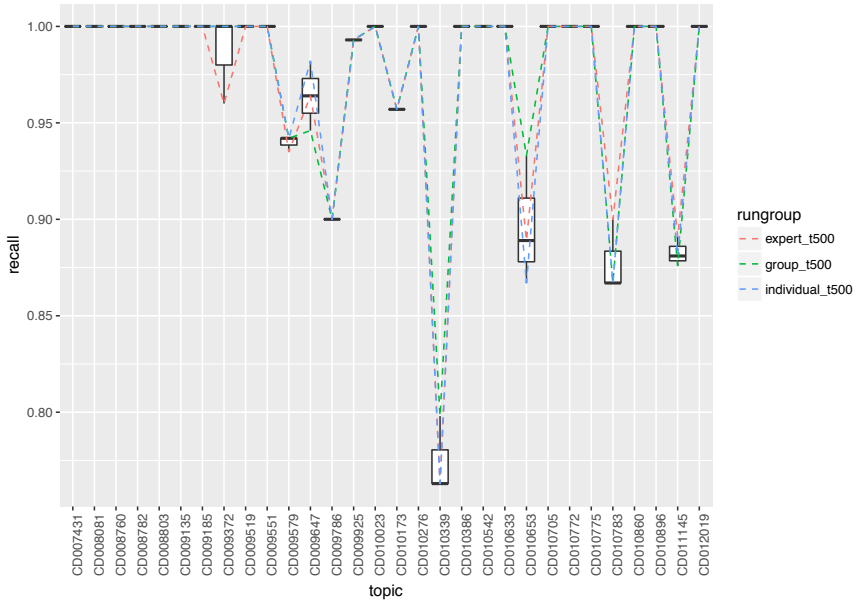



Fig. 3: Recall per topic for runs with $t = 500$ and $p = 50$.

the three approaches are practically indistinguishable given the same number of documents shown to the physician. We performed a Wilcoxon paired signed test for every pair of types of runs with $p = 50\%$ and $t = 500\%$, as well as $p = 50\%$ and $t = 1000\%$. The result of the statistical test confirms that there is no statistically significant difference among the performances of the runs.

On the other hand, for $p = 10\%$ there is a noticeable advantage of the individual query variants over the expert query variants. This is surprising to some extent, since it shows that the students were able to rewrite the information need better than the linguistic experts. We may explain this behaviour because while the two experts worked on all the 30 topics, the students worked on a single topic both individually and in groups. In this sense, the possibility to focus on a single topic may have allowed for a more in-depth domain research and terminological analysis. Furthermore, the fact of having worked in three people on the same topic may have helped to bring out linguistic aspects that have gone unnoticed by the two experts. Taking a timely estimate, the experts took about a month to complete the 30 topics, while the students had about two weeks available for both group and individual reformulation. All these factors may have influenced the query reformulation and consequently the effectiveness of the performance.

**Low Recall Topics** We perform a failure analysis on those topics for which the system did not achieve a recall of 100%. Since for $t = 1000$ we obtain an almost

perfect recall and there are no noticeable differences among the three types of runs, we decided to investigate the runs with threshold $t = 500$ and $p = 50\%$ since they achieve a good balance between recall, close to 0.95, and number of documents shown, around 40,000. In Figure 3, the box-plot summarizes the values of recall of the three runs for each topic, while the coloured lines highlight the (possible) differences among the three types of runs. There are only 10 topics that do not achieve a perfect recall. Among these topics, we focus topic CD010653 since it is the one with the largest difference in performance 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: 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.

## 5 Conclusions

In this paper, we presented a methodology for replicating an experiment of query rewriting in the context of medical eHealth. Following the approach by [14], we devised an active learning strategy that combines the information need and the boolean query of the physician with a ranked list of documents organized by the search engine in a continuous active learning framework which involved non-experts of the field of medicine.

The experiment consisted in re-proposing a task previously performed by a team of researchers for the CLEF eHealth Task 2: "Technologically Assisted Reviews in Empirical Medicine". Our working methodology was based on the analysis of linguistic and terminological aspects functional for the query rewriting in order to produce two variants of the same information need. The participants of this experiment were the students of the Master's Degree course in Modern Languages for International Communication and Cooperation of the University of Padua. They were required to rewrite the initial information need retrieve all the relevant documents for medical specific domains through the reformulation of an initial query given by an expert.

Experimental results showed that our approach allowed the TAR system to achieve a perfect recall on almost all the topics of the task with few significantly less documents compared to other CLEF participants of the same task. In terms of costs, the experts took about a month to complete the 30 topics, which means one day of work per topic, while the students had about two weeks available for both group and individual reformulation.

## References


1. Andrea L. Berez-Kroeker, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice, and Anthony C. Woodbury. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 561(1):1–18, 2017.
2. António Branco, Kevin Bretonnel Cohen, Piek Vossen, Nancy Ide, and Nicoletta Calzolari. Replicability and reproducibility of research results for human language technology: introducing an lre special section. *Language Resources and Evaluation*, 51(1):1–5, 2017.
3. Kevin B Cohen, Jingbo Xia, Christophe Roeder, and Lawrence Hunter. Reproducibility in natural language processing: A case study of two r libraries for mining pubmed/medline. In *In LREC 4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*, pages 6 – 12. European Language Resources Association (ELRA), 2016.
4. Gordon V. Cormack and Maura R. Grossman. Scalability of continuous active learning for reliable high-recall text classification. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1039–1048, New York, NY, USA, 2016. ACM.
5. Gordon V. Cormack and Maura R. Grossman. Technology-assisted review in empirical medicine: Waterloo participation in CLEF ehealth 2017. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017.
6. A.P. Cowie. *Phraseology: Theory, Analysis, and Applications*. Oxford Studies in Lexicography and Lexicology. OUP Oxford, 1998.
7. Giorgio Maria Di Nunzio, Federica Beghini, Federica Vezzani, and Geneviève 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 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017.
8. Giorgio Maria Di Nunzio, Federica Beghini, Federica Vezzani, and Geneviève Henrot. A reproducible approach with R markdown to automatic classification of medical certificates in french. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017.*, 2017.
9. Nicola Ferro. Reproducibility challenges in information retrieval evaluation. *J. Data and Information Quality*, 8(2):8:1–8:4, January 2017.
10. Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors. *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, volume 9626 of *Lecture Notes in Computer Science*. Springer, 2016.
11. Christopher Gandrud. *Reproducible Research with R and R Studio*. Chapman and Hall/CRC, second ed. edition, 2015.
12. D. Gouadec. *Terminologie: constitution des données*. AFNOR gestion. AFNOR, 1990.
13. Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker, editors. *CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview. In Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin,*

*Ireland, September 11-14, 2017.*, CEUR Workshop Proceedings. CEUR-WS.org, 2017.
14. Sarvnaz Karimi, Stefan Pohl, Falk Scholer, Lawrence Cavedon, and Justin Zobel. Boolean versus ranked querying for biomedical systematic reviews. *BMC Medical Informatics and Decision Making*, 10:58–58, 2010.
15. Donald E. Knuth. Literate programming. *Comput. J.*, 27(2):97–111, May 1984.
16. Marie-Claude L'Homme. Sur la notion de "terme". *Meta*, 50(4):1112–1132, 2005.
17. Mark Liberman. Validation of results in linguistic science and technology: Terminology, problems, and solutions. In António Branco, Nicoletta Calzolari, and Khalid Choukri, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA).
18. Antonio San Martín and Marie-Claude L'Homme. Definition patterns for predicative terms in specialized lexical resources. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 3748–3755, 2014.
19. Makoto Miwa, James Thomas, Alison O'Mara-Eves, and Sophia Ananiadou. Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics*, 51:242 – 253, 2014.
20. Giorgio Maria Di Nunzio. A study of an automatic stopping strategy for technologically assisted medical reviews. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 672–677, 2018.
21. F. Rastier. *Sémantique interprétative*. Formes sémiotiques. Presses universitaires de France, 1987.
22. Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, pages 109–126, 1994.
23. Gaurav Singh, James Thomas, and John Shawe-Taylor. Improving active learning in systematic reviews. *CoRR*, abs/1801.09496, 2018.
24. Federica Vezzani, Giorgio Maria Di Nunzio, and Geneviève Henrot. Trimed: A multilingual terminological database. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation LREC 2018, Miyazaky, Japan, May 7-12, 2018.*, 2018. In press.