Machine Learning in Information Retrieval - Classification of Precision Medicine Documents

Jakub Dutkiewicz and Czesław Jedrzejek

Politechnika Poznańska, Plac Marii Skłodowskiej-Curie 5, 60-965 Poznań, Poland
jakub.dutkiewicz@put.poznan.pl
czeslaw.jedrzejek@put.poznan.pl

Abstract. This work is concerned with machine learning in Information Retrieval when features are not explicit and the relevance assessment process indicates that term frequency methods should be complemented with classification procedures. We apply our approach to the Text REtrieval Conference (TREC 2017) PM Track[1]. We concentrate on so-called Human PM criterion. The goal of this study is to present a method, which automatically obtains value of this feature. We present the effectiveness of a simple boolean search based method and a method of converting a semi-structured document into a vectorized form. The vectorized document form is then used by various machine learning methods in order to solve this task. We achieve an accuracy of 77.89% with the use of a Support Vector Classifier.

Keywords: Precision Medicine · Classification · Information Retrieval.

1 Introduction

Precision medicine is a medical model that separates people into different groups for which medical decisions, practices, and interventions are addressed. The amount of knowledge required by a physician to put the findings of precision medicine into practice is huge and the goal of information retrieval is to help the best possible treatment for a particular patient. The Precision Medicine (PM) Track of the Text REtrieval Conference (TREC) deals with providing detailed information necessary for PM care. There are two target document collections for the Precision Medicine track: scientific abstracts and clinical trials. Here, we are concerned only with clinical trials. The corpus is derived from ClinicalTrials.gov, a repository of past, present, and future clinical trials in the U.S. and abroad. A total of 241,006 clinical trial descriptions compose the corpus provided to participants.

The TREC PM 2017 was the first one in the series for which there were special rules for evaluation[2]. The result assessment starts with determination of the so-called *Human PM* feature: The clinical trial (1) relates to humans, (2) involves

[1] https://trec.nist.gov
[2] http://www.trec-cds.org/relevance_guidelines.pdf

some form of cancer, (3) focuses on treatment, prevention, or prognosis of cancer, and (4) relates in some way to at least one of the genes in the topic[1]. In this study, we focus on determining the value of that feature using Boolean search and machine learning methods. This is one of the key features in determining the final relevance of the document. It is used by annotators in the process of assigning the relevance of a document. This study focuses solely on determining this feature value, not on the impact of the feature on the final ranking.

2 Related Work

In the original TREC PM 2017 track, the best overall results for clinical trials were achieved by the UD_GU_BioTM and UTDHLTRI teams. UTDHLTRI [2] used the Precision Medicine Drug Graph (PMDG) system, for both Topic Analysis and Topic Expansion. The works of UTDHLTRI are similar to this work, as the authors of this work use external resources in order to find synonymic forms of the query terms and thus expand the queries. A similar Boolean search, which is tested in our study, was applied by [3] for TREC PM 2021. Works of [4] make use of the task of classifying documents as either relevant or irrelevant to clinical studies. However, they put focus on scientific abstracts, instead of clinical trials. Clinical trials and scientific abstracs are quite different systems.

3 Method Overview

The classical Information Retrieval process is defined as a regression task. Given a set of queries and a corpus of documents, the task of the IR process is to assign a relevance score for each document-query pair. A ranking list is built upon the score assignments. Documents with a higher score are ranked higher than documents with a lower score. Contrarily to the regression-based Information Retrieval, the classification-based Information Retrieval task is to assign the relevance value of a document to each query. Relevance values are defined as a finite and discrete set of literals, which correspond to classes in the classification process. The TREC Precision Medicine challenge, in our opinion, is more fit to the classification-based Information Retrieval task. This is due to the well-defined relevance assessment process, which is conducted by human annotators.

3.1 Document Structure

Each processed document describes a clinical trial and is structured as a shallow tree. A root of the tree indicates that the document is in fact a clinical trial, while the leaves relate to specific fields of the trial, such as the title of the trial, its description, eligibility criteria, outcome of the clinical trial, etc. We categorize fields it is comprised of as either informative or non-informative. We do not perform any calculations on the non-informative fields. The informative fields include “brief title”, “official title”, “keywords”, “condition”, “summary”,

"description", "intervention name", "eligibility criteria", "primary outcome", and "secondary outcome".

Another piece of information comes in a form of a topic. Each topic consists of four fields: "disease", "gene", "demographics" and "other". In the process of determining the Human PM feature of the document, only the "gene" field is relevant.

For each informative field, we create a dedicated processing function. The task of this function is to get to the informational portion of a field, tokenize the text value and return a list of tokens, of which the field is comprised. We preserve the information of the origin of the token by adding an adequate prefix to the token.

We first use processing functions to create a vector space for document vectors. Each unique token within the fields corresponds to a dimension within the space. We then use the processing functions in order to create a set of vectors for documents. If a document contains a word, which corresponds to a vector component, the value of that component is set to 1, it is set to 0 otherwise.

4 Conducted Experiments

We use an annotated sample of a Clinical Trials corpus. The entire sample consists of 13019 documents, out of which 3959 documents are annotated as *Human PM*. We use all of the positively annotated documents. We randomly pick 3959 documents annotated as *Not PM* in order to complement the dataset.

Table 1. Evaluation of expansion methods for the Boolean Search classification.

Method	Accuracy	Precision	Recall
No expansion	66.25%	46.37%	76.98%
Gene name synonyms	69.71%	63.52%	72.50%
Targetted drugs	68.50%	57.89%	73.49%
Drugs and genes	69.44%	68.00%	69.45%

We test the effectiveness of a Boolean Search for the classification. We create a set of dedicated queries, which consist of terms, which correspond four aspects of precision medicine. We expand the queries with use of the external tools, such as MeSH database [3]. We present the effectiveness of Boolean Search in Tab. 1.

We employ a set of classical machine learning methods in order to solve this task. In the implementation, we use the sklearn [5] library. We test the expansion methods described in the previous section. We use vectorized documents as an input and the PM class as an output of the classification procedure. Detailed information about the conducted experiment can be found in Tab. 2. It should be noted that used evaluation measures are dedicated for determining the quality

[3] https://www.ncbi.nlm.nih.gov/mesh/

Table 2. Accuracy of machine learning based classification methods.

Method	No expansion	Gene expansion	Gene and drug expansion
Boolean search	66.25%	69.71%	69.44%
k-NN	70.11%	70.43%	70.62%
Random Forests	75.31%	74.74%	75.34%
Decision Tree	71.11%	70.51%	70.95%
Gaussian Naive Bayes	68.48%	68.34%	68.23%
Multinomial Naive Bayes	76.24%	75.90%	75.76%
SVM	77.84 %	77.43%	77.34%

of a classification system and are not used to evaluate the quality of the final ranking.

5 Conclusions

The following study proves that machine learning methods do outperform query-based Boolean search methods in the task of determining the Precision Medicine aspect of a clinical trial. When designing a method of information retrieval and evaluating results, one is facing very many options of taking into account document tags, informative features, synonyms, and relations. One has to go through numerous combinations of these and the effects of these options on the overall results get blurred. Therefore, here we concentrate on one aggregate feature: Human PM. Surprisingly, relating genes to drugs did not improve results.

References

1. Roberts, K., Demner-Fushman, D., Voorhees, E.M., Hersh W.R., Bedrick S., Lazar, A.J., Pant, S.: Overview of the TREC 2017 Precision Medicine Track. In: Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA (2017)
2. Goodwin, T.R., Skinner, M.A., Harabagiu, S.M. : Will Sorafenib Help?: Treatment-aware Reranking in Precision Medicine Search. In: Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA (2017)
3. Rybinski, M., Karimi, S. : UTD HLTRI at TREC 2017: Precision Medicine Track In: CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia (2021)
4. Ševa J, Wiegandt DL, Götze J, Lamping M, Rieke D, Schäfer R, Jähnichen P, Kittner M, Pallarz S, Starlinger J, Keilholz U, Leser U. VIST - a Variant-Information Search Tool for precision oncology. BMC Bioinformatics (2019)
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research vol.12 2825–2830 (2011)