Journal of Biomedical Informatics 46 (2013) 1145–1151

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Unsupervised mining of frequent tags for clinical eligibility text indexing

Riccardo Miotto [a], Chunhua Weng [a,b,*]

[a] Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA
[b] The Irving Institute for Clinical and Translational Research, Columbia University, New York, NY 10032, USA

ARTICLE INFO

Article history:
Received 27 February 2013
Accepted 29 August 2013
Available online 10 September 2013

Keywords:
Information storage and retrieval
Clinical trials
Tags
Information filtering
Eligibility criteria
Controlled vocabulary

ABSTRACT

Clinical text, such as clinical trial eligibility criteria, is largely underused in state-of-the-art medical search engines due to difficulties of accurate parsing. This paper proposes a novel methodology to derive a semantic index for clinical eligibility documents based on a controlled vocabulary of frequent tags, which are automatically mined from the text. We applied this method to eligibility criteria on ClinicalTrials.gov and report that frequent tags (1) define an effective and efficient index of clinical trials and (2) are unlikely to grow radically when the repository increases. We proposed to apply the semantic index to filter clinical trial search results and we concluded that frequent tags reduce the result space more efficiently than an uncontrolled set of UMLS concepts. Overall, unsupervised mining of frequent tags from clinical text leads to an effective semantic index for the clinical eligibility documents and promotes their computational reuse.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Parsing clinical eligibility text is important to leverage the reuse of clinical information for automatic decision support [1,2]. Following this assumption, various methods and techniques have been recently developed to transform clinical trial protocol text into computable representations that can benefit automated tasks (e.g., classification, clustering, discovery [3–9]). Several efforts specifically focused on clinical trial eligibility criteria, which define the characteristics that a research volunteer must possess to qualify for a clinical trial study. Some of these techniques index eligibility criteria using template-based semantic patterns or formal ontologies (e.g., [10,11]), whereas others extract from the text terms covered by the Unified Medical Language System (UMLS) lexicon [12] (e.g., [13]). Nevertheless, eligibility criteria generally remain as free text and underused in modern computational tasks such as search. As an example, ClinicalTrials.gov [14] does not process the eligibility criteria text when ranking trials in response to user queries [15]. A major reason is that a standardized and widely accepted parser for clinical trial eligibility criteria is not yet defined [16,17].

The indexing methods proposed in the literature generally parse each clinical trial separately without considering textual similarities among them. This results in an ever-expanding index with high dimensionality and high likelihood of presenting too specific, redundant, or irrelevant concepts for individual documents, which is not amenable for automated processing. To address this issue, we propose an alternative approach based on cross-processing eligibility criteria from multiple studies to mine a finite vocabulary of tags frequently shared by these trials, thus serving as a semantic index for eligibility text.[1]

In the information retrieval literature, the problem of document indexing and tagging has been robustly studied in different application scenarios as well as in terms of information theory [18–21]. Tags are generally used in exploratory retrieval, in which users engage in iterative cycles of document refinement and exploration of new information (as opposed to standard free-text retrieval). A controlled vocabulary of tags defines an interpretative layer of semantics over the text and its parsed representation, and generally leads to more effective retrieval than uncontrolled annotations [22]. For example, the use of a controlled vocabulary benefited different text search applications (e.g., [23–26]) as well as multimedia retrieval (e.g., [27,28]).

We hypothesize that (1) an unsupervised, fully automated data mining approach applied to the clinical trial repository can produce a finite set of tags that is frequently shared among all trials and (2) these frequent tags can lead to a general and stable index of eligibility text, which can leverage the automated processing of clinical trials. This method is potentially superior to other approaches by balancing and minimizing the sparseness of the index and increasing efficiency and specificity of retrieval operations [29]. Moreover, because of their high frequency, tags extracted

* Corresponding author. Address: Department of Biomedical Informatics, Columbia University, 622 W 168th Street, VC-5, New York, NY 10032, USA. Fax: +1 212 305 3302.

E-mail address: cw2384@columbia.edu (C. Weng).

[1] In this domain, tags are defined as meaningful multi-word semantic concepts automatically extracted from the text, such as, e.g., “breast carcinoma”, “diabetes”, “active malignancy”.

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.08.012

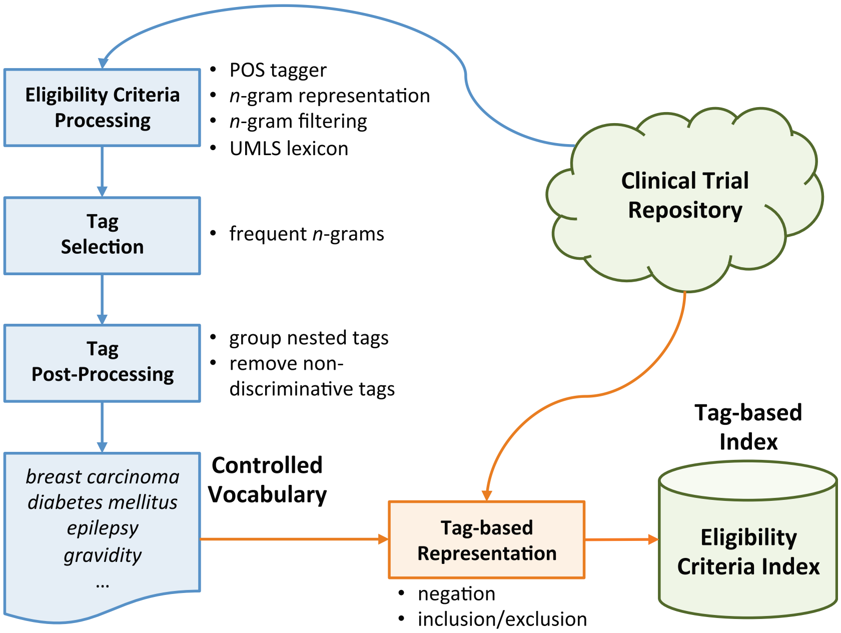

Fig. 1. Overview of tag mining and clinical trial eligibility criteria indexing.

from the text are likely to be more generic and intuitive than independent annotations, and therefore might be also more effective in helping users with interactive tasks [30].

The original contribution of this article is three-fold: we (1) present a novel method for mining a controlled vocabulary of frequent tags from clinical eligibility text; (2) apply this method to ClinicalTrials.gov and report statistics on tag distributions; and (3) propose to evaluate the effectiveness of frequent tags at filtering clinical trial search results when there is no gold standard available for comparing these tags against.

2. Material and methods

Fig. 1 depicts the proposed tag mining approach. First, the eligibility criteria in the repository are processed to extract potential tags individually. Then, only the tags meeting the frequency threshold are retained, post-processed, and used to index the trials.

2.1. Mining frequent tags

Eligibility text is often divided into two sections, one specifying whom to include (inclusion criteria) and the other whom to exclude (exclusion criteria). It might be listed explicitly and separately in a tabular format or expressed together as a general and vague text. Fig. 2 provides two examples of eligibility criteria with different structures, one tabular and the other free text. Classifying a criterion as inclusion or exclusion is straightforward when the division is explicitly reported, but more difficult when that division is implicit, since inclusion and exclusion criteria can be expressed in different ways. However, because tags are meant to identify high-level general concepts shared among clinical trials, the mining process can just focus on extracting tags regardless of their classification.

2.1.1. Eligibility text processing

The algorithm to process the eligibility criteria of a clinical trial relies on basic text processing techniques [31]. First, each criterion is automatically annotated with a part-of-speech (POS) tagger, defined in the *Natural Language Toolkit* [32], to identify the grammatical role of each word. In this application, the grammatical role of a word will be used only for noise reduction (e.g., to remove tags composed by only, e.g., verbs, adverbs); for this reason, we favored a general well-established solution rather than a more domain-related one [33]. The text is then processed to remove special characters and punctuation and to build all the possible *n*-grams (i.e., continuous sub-sequences of *n* words).[2] *N*-grams composed of only English stop words or irrelevant grammatical structures are removed.

Lastly, each *n*-gram is matched against the UMLS Metathesaurus and retained only if at least one substring of it is a recognizable UMLS concept. Moreover, we considered only those UMLS concepts appearing in semantic categories most relevant to the clinical trial domain [34] (i.e., 27 different semantic types out of 136, including, e.g., “Disease or Syndrome”, “Individual Behavior”, “Finding”) in order to reduce the number of extraneous tags. As an example, “malignancy within the past 5 years” is considered a valid *n*-gram because at least one word, “malignancy”, is present in the part of the UMLS lexicon considered, even if the entire sentence is not.[3] Each *n*-gram term found in the UMLS lexicon is also normalized according to its preferred Concept Unique Identifier (CUI) in order to reduce the sparseness of the concepts. Using the CUIs also enables the handling of synonyms, since similar concepts are aligned to the same preferred term because of the UMLS specification (e.g., “atrial fibrillation” and “auricular fibrillation” are both mapped to “atrial fibrillation”). This allows defining a vocabulary possibly composed by semantically unique tags. After this process, each clinical trial’s eligibility criteria are summarized by a set of UMLS CUI-based *n*-grams representing the criteria’s relevant concepts.

2.1.2. Frequent tag selection

Given a repository of clinical trials and their *n*-gram-based representations, the set of tags is obtained by retaining the *n*-grams

[2] We used *n*-grams with lengths ranging from 1 to 10. In fact, we observed in preliminary experiments that tags longer than 7 words were unlikely to appear frequently. Therefore, we used 10 as maximum length to handle potential outliers.

[3] If the regular UMLS lexicon were used, the previous example “malignancy within the past 5 years” would have had three terms correctly matched (i.e., “malignancy”, “past”, and “years”).

Download English Version:

https://daneshyari.com/en/article/10355475

Download Persian Version:

https://daneshyari.com/article/10355475

Daneshyari.com