Towards a Mixed Approach to Extract Biomedical Terms from Text Corpus

Juan Antonio Lossio Ventura, LIRMM, University Montpellier 2, Montpellier, France & CNRS, Paris, France
Clement Jonquet, LIRMM, University Montpellier 2, Montpellier, France & CNRS, Paris, France
Mathieu Roche, UMR TETIS, Cirad, Irstea, AgroParisTech, Montpellier, France
Maguelonne Teisseire, UMR TETIS, Cirad, Irstea, AgroParisTech, Montpellier, France

ABSTRACT

The objective of this paper is to present a methodology to extract and rank automatically biomedical terms from free text. The authors present new extraction methods taking into account linguistic patterns specialized for the biomedical domain, statistic term extraction measures such as C-value and statistic keyword extraction measures such as Okapi BM25, and TFIDF. These measures are combined in order to improve the extraction process and the authors investigate which combinations are the more relevant associated to different contexts. Experimental results show that an appropriate harmonic mean of C-value associated to keyword extraction measures offers better precision, both for single-word and multi-words term extraction. Experiments describe the extraction of English and French biomedical terms from a corpus of laboratory tests available online. The results are validated by using UMLS (in English) and only MeSH (in French) as reference dictionary.

Keywords: Biomedical Natural Language Processing (BioNLP), Biomedical Term Extraction, Biomedical Terminologies and Ontologies, Biomedical Thesaurus, Statistic Measure, Text Mining

INTRODUCTION

The huge amount of data available online today is often composed of plain text field, for instance, clinical trial descriptions, adverse event reports or electronic health records. These texts often contain the real language (expressions and terms) used by the community. Although in the biomedical domain there exist hundred of terminologies and ontologies to describe such languages (Noy et al., 2009), those terminologies often miss concepts or possible alternative terms for those concepts. Our motivation is to improve the precision of automatic terms extraction process, the main reason for this, is that language evolves faster than our ability to formalize and catalog it. This is even more true for French in which the number of terms
formalized in terminologies is significantly less important than in English.

NLP (natural language processing) tools and methods enable to enrich biomedical dictionaries from texts. Automatic Term Recognition (ATR) is an approach in language technology that involves the extraction of technical terms from domain-specific language corpora (Zhang et al., 2008). In addition, Automatic Keyword Extraction (AKE) is the process of extracting the most relevant words or phrases in a document. Keywords, which we define as a sequence of one or more words, provide a compact representation of a document’s content. Two popular AKE measures are Okapi $BM25$ and $TFIDF$, also called weighting measures. These two fields are summarized in Table 1.

In our work, we adopt as baseline measures an ATR method, $C$-value (Frantzi et al., 2000), and the best two AKE methods (Hussey et al., 2012). Indeed, the $C$-value, compared to other ATR methods, often gets best precision results and especially in biomedical studies (Knoth et al., 2009; Zhang et al., 2008; Zhang et al., 2004). Moreover, this measure is defined for multi-word term extraction but can be easily adapted for single-word term (presented later on) and it has never been applied to French text, which is appealing in our case. Okapi and $TFIDF$ are the best AKE methods (Hussey et al., 2012). We propose to define new extraction methods by combining in different manners ATR and AKE measures, in order to rank the best candidate terms. Our experiment results underline the precision efficiency gain with the proposed methods. We give priority to precision in order to focus on extraction of new valid terms (precision) rather than on missed terms (recall), i.e., for a candidate term to be a valid biomedical term or not.

The rest of the paper is organized as follows: section “Related Work” describes the state of the art in the field of ATR, and specially the methods based on $C$-value; section “Proposed Approach” presents our proposal of ranking measures; section “Experiments and Results” details and discusses the conducted experiments and the associated results; and section “Conclusion” concludes the paper.

**RELATED WORK**

ATR studies can be divided into four main categories: (i) rule-based approaches, (ii) dictionary based approaches, (iii) statistical approaches, and (iv) hybrid approaches. Rule-based approaches for instance (Gaizauskas et al., 2000), attempt to recover terms thanks to the formation patterns, the main idea is to build rules in order to describe naming structures for different classes using orthographic, lexical, or morphosyntactic characteristics. Dictionary-based approaches use existing terminology resources in order to locate term occurrences in texts (Krauthammer et al., 2004). Statistical approaches are often built for extracting general terms (Eck et al., 2010). The most basic measure is frequency. $C/NC$-value (Frantzi et al., 2000) is another statistical method well known in the literature that combines statistical and linguistic information for the extraction of multi-word and nested terms. While most studies address specific types of entities, $C/NC$-value is a domain-independent method, used for extracting terms

<table>
<thead>
<tr>
<th>Table 1. Differences between ATR and AKE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong></td>
</tr>
<tr>
<td><strong>Output</strong></td>
</tr>
<tr>
<td><strong>Domain</strong></td>
</tr>
<tr>
<td><strong>Exemples</strong></td>
</tr>
</tbody>
</table>
Related Content

Data Mining and Meta-Analysis on DNA Microarray Data
www.irma-international.org/article/data-mining-meta-analysis-dna/70016/

In Silico Analysis of the CST6 Tumor Suppressor Gene
www.irma-international.org/article/in-silico-analysis-of-the-cst6-tumor-suppressor-gene/97741/

Efficient Mining Frequent Closed Discriminative Biclusters by Sample-Growth: The FDCluster Approach
Miao Wang, Xuequn Shang, Shaohua Zhang and Zhanhuai Li (2012). *Computational Knowledge Discovery for Bioinformatics Research* (pp. 84-103).
www.irma-international.org/chapter/efficient-mining-frequent-closed-discriminative/66706/

DNA Sequence Visualization
www.irma-international.org/chapter/dna-sequence-visualization/4246/

Methods for the Evaluation of Right Ventricular Volume Using Ultrasound on a Catheter, in Intensive Care Unit
www.irma-international.org/article/methods-evaluation-right-ventricular-volume/78391/