BIOTEX: A system for Biomedical Terminology Extraction, Ranking, and Validation

Juan Antonio Lossio-Ventura[1], Clement Jonquet[1],
Mathieu Roche[1,2], and Maguelonne Teisseire[1,2]

[1]University of Montpellier 2, LIRMM, CNRS - Montpellier, France
[2]Irstea, CIRAD, TETIS - Montpellier, France
juan.lossio@lirmm.fr, jonquet@lirmm.fr,
mathieu.roche@cirad.fr, maguelonne.teisseire@teledetection.fr

Abstract. Term extraction is an essential task in domain knowledge acquisition. Although hundreds of terminologies and ontologies exist in the biomedical domain, the language evolves faster than our ability to formalize and catalog it. We may be interested in the terms and words explicitly used in our corpus in order to index or mine this corpus or just to enrich currently available terminologies and ontologies. Automatic term recognition and keyword extraction measures are widely used in biomedical text mining applications. We present BIOTEX, a Web application that implements state-of-the-art measures for automatic extraction of biomedical terms from free text in English and French.

1 Introduction

Within a corpus, there is different information to represent, with different communities to express that information. Therefore, the terminology and vocabulary is often very corpus specific and not explicitly defined. For instance in medical world, terms employed by lay users on a forum will necessarily differ from the vocabulary used by doctors in electronic health records. We thus intend to offer users an opportunity to automatically extract biomedical terms and use them for any natural language, indexing, knowledge extraction, or annotation purpose. Extracted terms can also be used to enrich biomedical ontologies or terminologies by offering new terms or synonyms to attach to existing defined classes. Automatic Term Extraction (ATE) methods are designed to automatically extract relevant terms from a given corpus.[1]. Relevant terms are useful to gain further insight into the conceptual structure of a domain. In the biomedical domain, there is a substantial difference between existing resources (ontologies) in English and French. In English there are about 7 000 000 terms associated with about 6 000 000 concepts such as those in UMLS or BioPortal [7]. Whereas, in French there are only about 330 000 terms associated with about 160 000 concepts [6]. French ontologies therefore have to be populated and tool like BIOTEXwill help for this task. Our project involves two main stages: (i) Biomedical term extraction, and (ii) Ontology population, in order to populate ontologies with the extracted terms.

[1] We refer to ATE when terms extracted are not previously defined in existing standard ontologies or terminologies. We refer to 'semantic annotation' when term extracted can be attached or match to an existing class (URI) such as in [8]. Both approaches are related to Named Entity Recognition (NER), which automatically extracts name of entities (disease, person, city).

In this paper, we present BIOTEX, an application that performs the first step. Given a text corpus, it extracts and ranks biomedical terms according to the selected state-of-the-art extraction measure. In addition, BIOTEX automatically validates terms that already exist in UMLS/MeSH-fr terminologies. We have presented different measures and performed comparative assessments in other publications [4, 5]. In this paper, we focus on the presentation of BIOTEX and the use cases it supports.

2 Related work and available extraction measures

Term extraction techniques can be divided into four broad categories: (i) **Linguistic approaches** attempt to recover terms via linguistic patterns [3]. (ii) **Statistical methods** focus on external evidence through contextual information. Similar methods, called Automatic Keyword Extraction (AKE), are geared towards extracting the most relevant words or phrases in a document. These measures, such as *Okapi BM25* and *TF-IDF*, can be used to automatically extract biomedical terms, as we proposed in [4]. These two measures are included in BIOTEX. (iii) **Machine Learning** is often designed for specific entity classes and thus integrate term extraction and term classification. (iv) **Hybrid methods**. Most approaches combine several methods (typically linguistic and statistically based) for the term extraction task. This is the case of *C-value* [2], a very popular measure specialized in multi-word and nested term extraction.

In [4], we proposed the new hybrid measures *F-TFIDF-C* and *F-OCapi*, which combine *C-value* with *TF-IDF* and *Okapi* respectively to extract terms and obtain better results than *C-value*. In [5], we propose *LIDF-value* measure based on linguistic and statistical information. We offer all of these measures within BIOTEX. Our measures were evaluated in terms of *precision* [4, 5] and obtained the best results over the top k extracted terms ($P@k$) on several corpora (LabTestOnline, GENIA, PubMed). For instance, on a GENIA corpus, *LIDF-value* achieved 82% for $P@100$, thus improving the *C-value* precision by 13%, and 66% for $P@2000$, with an improvement of 11%. BIOTEX allows users to assess the performances of measures with different corpora.

A detailed study of related work revealed that most existing systems implementing statistical methods are made to extract keywords and, to a lesser extent, to extract terminology from a text corpus. Indeed, most systems take a single text document as input, not a set of documents (as corpus), for which the *IDF* can be computed. Most systems are available only in English. Table 1 shows a quick comparison with *TerMine* (*C-value*), the most commonly used application, and *FlexiTerm*, the most recent one.

Table 1. Brief comparison of biomedical terminology extraction applications.

	$BioTex$	$TerMine$	$FlexiTerm$
Languages	en/fr	en	en
Type of Application	Desktop/Web	Web	Desktop
License	Open	Open	Open
Processing Capacity	No Limits / $<$ 6 MB	$<$ 2 MB	No Limits
Possibility to save results	XML	-	CSV
POS tool	TreeTagger	Genia/TreeTagger	Stanford POS
# of Implemented Measures	8	1	1

`http://www.nactem.ac.uk/software/termine/`
`http://users.cs.cf.ac.uk/I.Spasic/flexiterm/`

3 Implementation of BIOTEX

BIOTEX is an application for biomedical terminology extraction which offers several baselines and new measures to rank candidate terms for a given text corpus. BIOTEX can be used either as: (i) a Web application taking a text file as input, or (ii) as a Java library. When used as a Web application, it produces a file with a maximum of 1200 ranked candidate terms. Used as a Java library, it produces four files with ranked candidate terms found in the corpus, respectively, unigram, bigram, 3-gram and all the 4+ gram terms. BIOTEX supports two main use cases:

(1) **Term extraction and ranking measures:** As illustrated by the Web application interface, Figure 1 (1), BIOTEX users can customize the workflow by changing the following parameters:
   - Choose the corpus *language* (i.e., English or French), and the *Part-of-Speech* (PoS) tagger to apply. Note that we tested three POS-tagger tools but currently only TreeTagger is available within BIOTEX.
   - Select a number of *patterns* to filter out the candidate terms (200 by default). Those reference patterns (e.g., noun-noun, noun-prep-noun, etc.) were built with terms taken from UMLS for English and MeSH-fr for French. They are ranked by frequency.
   - Select the *type of terms* to extract: all terms (i.e., single- and multi-word terms) or multi-word terms only.
   - Select the *ranking measures* to apply.

(2) **Validation of candidate terms:** After the extraction process, BIOTEX automatically validates the extracted terms by using UMLS (Eng) & MeSH-fr (Fr). As illustrated in Figure 1 (2), these validated terms are displayed in green, specifying the used knowledge source and the others in red. Therefore, BIOTEX allows someone to easily distinguish the classes annotating the original corpus (in green) from the terms that maybe also considered relevant for their data, but need to be curated (in red). The last ones may be considered candidates for ontology enrichment.

4 Conclusions and Future Work

In this article, we present the BIOTEX application for biomedical terminology extraction. It is available for online testing and evaluation but can also be used in any program as a Java library (POS tagger not included). In contrast to other existing systems, this system allows us to analyze a French corpus, manually validate extracted terms and export the list of extracted terms. We hope that BIOTEX will be a valuable tool for the biomedical community. It is currently used in a couple of test-beds within the SIFR project (`http://www.lirmm.fr/sifr`). The application is available at `http://tubo.lirmm.fr/biotex/` along with a video demonstration `http://www.youtube.com/watch?v=EBbkZj7HcL8`. For our future validations, we will enrich our validation dictionaries with BioPortal [7] terms for English and CISMeF [9] terms for French. In the future, we will offer disambiguation features using the Web to find the context in order to populate biomedical ontologies with the new extracted terms (red terms), while looking into the possibility of extracting relations [1] between new terms and already known terms.

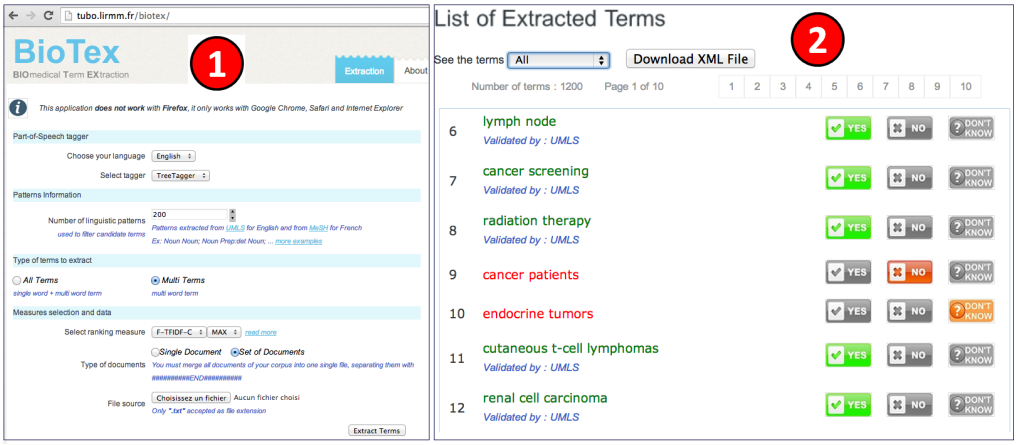

Fig. 1. (1) Interface: term extraction. (2) Interface: term validation. Users can export the results for offline processing.

Acknowledgments. This work was supported in part by the French National Research Agency under JCJC program, grant ANR-12-JS02-01001, as well as by University of Montpellier 2, CNRS, IBC of Montpellier project and the FINCyT program, Peru

References

1. Abacha, A. B., Zweigenbaum, P.: Automatic extraction of semantic relations between medical entities: a rule based approach. Journal of Biomedical Semantics, vol. 2 (2011)
2. Frantzi K., Ananiadou S., Mima, H.: Automatic recognition of multiword terms: the C-value/NC-value Method. International Journal on Digital Libraries, vol. 3, pp. 115-130, (2000)
3. Gaizauskas, R., Demetriou, G., Humphreys, K.: Term recognition, classification in biological science journal articles. Proceeding of the Computional Terminology for Medical, Biological Applications Workshop of the 2 nd International Conference on NLP, pp. 37-44 (2000)
4. Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire M.: Towards a Mixed Approach to Extract Biomedical Terms from Text Corpus. International Journal of Knowledge Discovery in Bioinformatics, IGI Global. vol. 4, pp. 1-15, Hershey, PA, USA (2014)
5. Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire M.: Yet another ranking function to automatic multi-word term extraction. Proceedings of the 9th International Conference on Natural Language Processing (PolTAL'14), Springer LNAI. Warsaw, Poland (2014)
6. Neveol, A., Grosjean, J., Darmoni, S., Zweigenbaum, P.: Language Resources for French in the Biomedical Domain. 9th International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland (2014)
7. Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M., Chute, C.G., Musen, M. A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research, vol. 37(suppl 2), pp 170–173. (2009)
8. Jonquet, C., Shah, N.H., Youn, C.H., Callendar, Chris, Storey, M-A, Musen, M.A.: NCBO Annotator: Semantic Annotation of Biomedical Data. 8th International Semantic Web Conference, Poster and Demonstration Session Washington DC, USA (2009)
9. Darmoni, S.J., Pereira, S., Sakji, S., Merabti, T., Prieur, E. Joubert, M., Thirion, B.: Multiple Terminologies in a Health Portal: Automatic Indexing and Information Retrieval. 12th Conference on Artificial Intelligence in Medicine, LNCS 5651, pp.255-259, Verona, Italy (2009)