Knowledge-Based Systems 57 (2014) 104–118

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Towards a Protein–Protein Interaction information extraction system: Recognizing named entities

Roxana Danger [a,*,1], Ferran Pla [b], Antonio Molina [b], Paolo Rosso [b]

[a] Dept. of Computing, Imperial College London, South Kensington Campus, UK
[b] Dpto. de Sistemas Informáticos y Computación, Universitat Politècnica de València, Spain

ARTICLE INFO

Article history:
Received 29 May 2013
Received in revised form 9 December 2013
Accepted 9 December 2013
Available online 17 December 2013

Keywords:
Biomedical named entity recognition
Protein–Protein Interaction
Dictionary look-up
Conditional random field
Support vector machine

ABSTRACT

The majority of biological functions of any living being are related to Protein–Protein Interactions (PPI). PPI discoveries are reported in form of research publications whose volume grows day after day. Consequently, automatic PPI information extraction systems are a pressing need for biologists. In this paper we are mainly concerned with the named entity detection module of PPIES (the PPI information extraction system we are implementing) which recognizes twelve entity types relevant in PPI context. It is composed of two sub-modules: a dictionary look-up with extensive normalization and acronym detection, and a Conditional Random Field classifier. The dictionary look-up module has been tested with Interaction Method Task (IMT), and it improves by approximately 10% the current solutions that do not use Machine Learning (ML). The second module has been used to create a classifier using the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA'04) data set. It does not use any external resources, or complex or ad hoc post-processing, and obtains 77.25%, 75.04% and 76.13 for precision, recall, and F1-measure, respectively, improving all previous results obtained for this data set.
© 2013 Elsevier B.V. All rights reserved.

1. Introduction

The study of Protein–Protein Interactions (PPI) has become crucial for many research topics in biology, since they are intrinsic to virtually every cellular process [1]. The majority of PPI information is available in the form of research articles whose volume grows day after day. In order to provide biologists with fast access to all this information, curators from various research institutes are dedicated to extracting the most important descriptions from publications, and to storing the extracted data on Protein Interaction Databases, such as: the Munich Information Center for Protein Sequence (MIPS) protein interaction Database [2]; the Biomolecular Interaction Network Database (BIND) [3]; the Database of Interacting Proteins (DIP) [4]; the Molecular Interaction Database (MINT) [5]; the protein Interaction Database (IntAct) [6]; the Biological General Repository for Interaction Datasets (BioGRID) [7]; and the Human Protein Reference Database (HPRD) [8].

Currently, the curation load is shared amongst all databases, and is built on the MIMIx [9] (Minimum Information about a Molecular Interaction Experiment) resources, part of the Proteomics Standards Initiative (PSI), of the Human Proteome Organization (HUPO).[2] The MIMIx resources are composed by the MIMIx guidelines, the PSI-MI XML interchange format, and the corresponding controlled vocabularies for molecular interaction description.

The curated data are regularly interchanged using the common standard PSI-MI extensible markup language (XML). However, expert curators may need a whole day to extract all the relevant information from an article,[3] and it is estimated that about 5% of Pubmed articles are referred to PPI.[4] Therefore, a semi-automatic processing of these papers is a pressing need for biologists and a challenge for bioinformatics researchers.

Automatic PPI information extraction involves many tasks: article classification (as positive/negative according to the PPI subject), biology named entity detection (especially for genes and proteins), normalization, and entity relation identification (especially interacting genes/proteins), which have been extensively discussed, mainly during the BIOCREATIVE Challenges.[5]

In this paper we introduce the general architecture of our system for automatizing the process of PPI information extraction, PPIES, as well as its module for *named entity detection*, and the

* Corresponding author. Tel.: +44 (0)207 594 8225.
E-mail address: rdanger@imperial.ac.uk (R. Danger).

[1] This work was developed while the first author was working for the ELiRF Research Group at the Department of Computer Systems and Computation, Universidad Politécnica de Valencia, Spain.

[2] http://www.psidev.info/.

[3] Based on answer to query 26 at http://biocreative.sourceforge.net/ppi_questions.html.

[4] Based on motivation for ACT-BC-III at http://www.biocreative.org/tasks/biocreative-iii/ppi/.

[5] http://www.biocreative.org.

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.12.010

results it obtains. The *named entity detection module* allows the complete set of entities described by MIMIx to be identified. It is a crucial step for the information extraction system and can also alleviate the curator's task, since all important detected entities can be highlighted, and the curator could go directly to extract the relevant information around them. It is composed of a dictionary look-up and a Conditional Random Field (CRF) classifier.

The dictionary look-up searches in a text for entities which can be associated to a relatively stable set of terms for *organisms, interaction detection and participant identification methods, interaction types, interactor types, biological roles*, and *tissue types*, using soft matching. To assess the performance of this module is a difficult task, as there are no available corpora in the PPI context tagged with all these entities. We have, however, used this module to solve the IMT task of BIOCREATIVE III [10], which consists in the recognition of the interaction detection methods used in PPI discovery.

The CRF classifier searches for entities that cannot be described through a dictionary, due to their incompleteness or inaccuracy (new molecules are discovered day after day, new synonyms and acronyms for a specific entity can be introduced and, depending on the data source, the list of names can be more or less complete and the ambiguity more or less difficult to resolve), as in the case of proteins, cell lines, cell types, DNA, and RNA molecules. In this sense the JNLPBA'04 corpus [11] is the only available resource containing biomedical texts tagged by these entity types.

In the following section a literature review related to our named entity detection module is presented. A general overview of the PPIES system as well as of the implementation details of the named entity detection module are given in Section 3. Section 4 describes and discusses the obtained results. Finally, in Section 5 conclusions are drawn and future work directions are discussed.

2. Background

The most important details related to the dictionary look-up systems are highlighted below in Section 2.1. The JNLPBA'04 corpus and the solutions described in the literature for the annotation of its entities are summarized in Section 2.2.

2.1. Dictionary look-up

Dictionary look-up, a type of string matching [12] algorithm, is useful in many Natural Language Processing applications, since it allows to retrieve terms of a given controlled vocabulary (CV) from a raw text. Normally, this vocabulary is formed by tuples of $(Id, term, entity\_type)$. The identifiers, *Id*, can be used to normalize the recognized *terms*, which are also linked to *entity types*. The accuracy of a dictionary look-up depends on the measure function that is used to compute the matching score level between texts and terms. Examples of soft matching measures are *n-gram similarity*, *Levenshtein distance* [13], and the *Jaro-Winkler* measure [14]. More sophisticated approaches combine different soft matching measures and/or learn the weights of their parameters from the dictionary (e.g. [15–18]).

Various techniques that optimize the time searching and the similarity measures have been proposed for dictionary look-up (e.g. [19–24]). Currently, search engines are used to create indexes of CV and/or of texts and allow retrieve texts associated to terms entered by users. Many bibliographic databases, e.g. PubMed, PubMed Central, Science Citation Index Expanded, ACM, Google Scholar, Citebase and Embase, uses such approach, but only a few of them uses a CV for indexing texts.

PubMed and Embase are the most important examples, in the biomedical area, using CV to index texts. Indexing texts with a CV implies that each text is processed by a dictionary look-up algorithm to capture the mentioned CV terms, and to maintain the recognized terms along with the texts in the index. Embase [23] indexes texts using their own Emtree thesaurus, formed by approximately 60,000 biomedical terms with a large coverage of chemicals and drug terminology. Part of the database is automatically indexed, but the details of the dictionary look-up algorithm are not provided.

PubMed is indexed using the NLM (National Library of Medicine[6]) Medical Text Indexer (MTI) which in turn uses MetaMap (see [21] for an overview), a dictionary look-up for UMLS Metathesaurus [25]. Other efforts for annotating texts for UMLS and MeSH are MicroMeSH [26], CHARTLINE [27] CLARIT [28], SAPHIRE [29], KnowledgeMap [30], MGREP [31].

MetaMap is the best well-known technology, in the biomedical field for dictionary look-up. It has merged in one tool all experiences for annotating biomedical texts and outperforms almost all other similar systems (an exception is KnowledgeMap in the context of biological process). Text processing in MetaMap is carried out using a series of linguistic steps for obtaining a mapping between segments of a text and concepts in UMLS: (1) tokenization, sentence boundary determination and acronym/abbreviation identification; (2) part-of-speech tagging; (3) lexical lookup of input words in the SPECIALIST lexicon; (4) a shallow parser to identify phrases and their lexical heads; (5) each phrase is analysed for obtaining different variations, and the Metathesaurus terms matching the input text, called candidates, are selected and evaluated; and (6) a mapping between text phrases and a combination of the candidates is generated and evaluated. The mapping is filtered, optionally disambiguated, and given as final result. It is out of the scope of this paper to describe the whole complexity behind each of these steps. The interested reader can refer to [21] for a deeper understanding.

Using MetaMap and adjusting it according to a particular use case is difficult. One the one hand, it is open-source but uses SICStus Prolog which is not-open source software. On the other hand, many parameters (e.g. the syntactic analysis algorithms and/or models) cannot be configured at the level granularity that a developer could desire. So, our goal is to construct a highly-configurable CV lookup system with similar linguistic approach as in MetaMap for terms in the context of PPI,[7] based only on open-source developments. The complete description of the system is given in Section 3.1.

As previously mentioned, the dictionary lookup module will be used to solve the IMT task of BIOCREATIVE III. IMT task consists in annotating full articles with the experimental methods that were used to detect a Protein–Protein Interaction (PPI), where the PSI-MI ontology is used to obtain the controlled vocabulary that characterizes the experimental methods. The data given by the organizers of the BIOCREATIVE III edition are summarized in Table 1. The task was evaluated considering macro and micro-observations, that is, considering only the documents for which a result was returned and considering all documents in the test set, respectively.

Eight teams participated in this task [10]. Six of them used ML approaches to perform the required task. Basically, they focused the task as a multi-label, multi-class classification problem at document or chunk level based on bag-of-words after a lexical analysis (a few teams used n-grams and named entity recognition). The probability output of the classifiers was used to rank and select the final list of experimental methods described in each article. Respect to the macro values, the system described in [32] obtained

[6] www.nlm.nih.gov.

[7] However, we have not yet addressed the word disambiguation problem.

Download English Version:

https://daneshyari.com/en/article/402713

Download Persian Version:

https://daneshyari.com/article/402713

Daneshyari.com