Information Sciences 360 (2016) 170–180

Contents lists available at ScienceDirect

Information Sciences

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

Question-driven topic-based extraction of Protein–Protein Interaction Methods from biomedical literature☆

John Atkinson[a,*], Gerardo Montecinos[a], Dorothy Curtis[b]

[a] Department of Computer Sciences, Universidad de Concepcion, Concepcion, Chile
[b] Computer Science and Artificial Intelligence Laboratory, MIT, USA

ARTICLE INFO

Article history:
Received 6 April 2015
Revised 28 March 2016
Accepted 2 April 2016
Available online 4 May 2016

Keywords:
Biomedical text mining
Topic models
Information access
Natural-language processing

ABSTRACT

This paper proposes a novel topic-based model for identifying experimental mentions of *Protein–Protein Interaction Method* (PPIM) in the biomedical literature. The model combines topic-based classification models and some basic question-answering extraction techniques aiming at effectively detecting and identifying PPIM mentions on *Protein–Protein Interactions*. Unlike other state-of-the-art approaches, the approach captures underlying relationships within both input and output concept spaces by assuming the extraction task to be strongly driven by context provided by experts, usually in the form of a question to guide the search. Results indicate our topic-based question-driven approach obtained better results than other unsupervised learning probabilistic latent space models for detecting correct answers (PPIM mentions).

1. Introduction

Traditionally, research on biomedical text mining has focused on extracting bio-entities (i.e., proteins) interactions from the biological literature [13]. However, little work has been done on the extraction of experimental *Protein–Protein Interaction Methods* (PPIM) that give account of theories or interactions between those bio-entities. It is crucial to annotate the detected PPIM as different PPIMs provide different degrees of reliability on the interactions. However, the diversity of the PPIMs mentioned in the literature makes the automatic extraction quite challenging. Manually annotating PPIM mentions is a time-consuming task: the curation of a manuscript may take up 3 h of an expert curator. Hence, there is great practical demand for automatically extracting PPIM mentions [16].

The diversity of PPIM mentions is the major obstacle for automatic PPIM extraction as multiple authors use different words and phrases to describe the same PPIMs. For example, the detection PPIM *"two hybrid"* has several related synonyms including *"2-hybrid", "classical two hybrid", "Gal4 transcription regeneration"*, etc. and one exact synonym (e.g. *"2 hybrid"*) in a specific-purpose ontology. String matching algorithms and dictionary-based methods have been used to capture these variations, however, results have shown low accuracy for identifying named PPIMs [3]. Furthermore, in the best case, techniques used to identify PPIM mentions are unable to associate them to the theories or interactions they refer to [14,17–19]. Recent

☆ This research was supported by FONDECYT-Chile under Grant number 1130035: *"An Evolutionary Computation Approach to Natural-Language Chunking for Biological Text Mining Applications"* and MIT-Chile Seed fund *"MIT-Chile Collaboration for Medical Text Summarization for Patient Literacy"*.

* Corresponding author. Tel.: +56 412204305.
E-mail addresses: atkinson@inf.udec.cl (J. Atkinson), dcurtis@csail.mit.edu (D. Curtis).

http://dx.doi.org/10.1016/j.ins.2016.04.039
0020-0255/

approaches view the extraction problem as a classification task in which automatic supervised classifiers are built to identify the existence of named PPIMs [2].

However, traditional discriminative classifiers fail to capture underlying relationships within both input and output concept spaces [8] as they are unaware of structural features of the named entities. On the other hand, real-life applications of PPIM mentions extraction are usually based on additional context provided by an expert's question to guide the search. Thus, the problem can be seen as expressing a question in such a way that an extraction technique is 'forced' to find a *PPIM* as an answer. These include factoid questions looking for direct answers such as *What experimental PPIM is used for X?*, *how-to* questions looking for procedure answers such as *How was X detected in Y?*, etc. Unlike traditional keywords-based matching strategies, question-answering techniques contribute some additional syntactic and semantic features which are valuable to find relevant answers.

Accordingly, in this paper a topic-based question-driven approach is proposed to extract PPIM mentions concerning *Protein–Protein Interactions* (PPI) from biomedical natural language texts. The model combines lexico-syntactical analysis and semantic grammar techniques [6], question-answering techniques and unsupervised learning to accurately detect experimental PPIM mentions. Thus, our claim is that combining natural language techniques, topic-based models and question-answering techniques can be more effective to capture underlying relationships which support the PPIM mentions detection task. It assumes the extraction task to be strongly context-driven, which usually exists in the form of a question to guide the search for suitable PPIMs within the literature.

Thus, the main contributions of the paper are twofold:

1. A novel adaptive question-driven extraction approach to effectively identify experimental PPIM mentions in the biomedical literature based on underlying lexico-syntactical knowledge and semantic relationships captured via question-answering techniques.
2. A model that combines topic-based unsupervised learning such as LDA and QA answer extraction techniques aiming at detecting and identifying PPIM mentions.

2. Related work

One of the key problems in several technical and scientific areas concerns the identification of PPIMs used by different theories when describing experimental research [16].

For instance, biomedical experts usually look for mentions of experimental PPIM used by theories describing PPI as it may be responsible for causing several biological phenomena. Hence automatically identifying the methods used in a research can provide insights on certain diseases, which in turn, lead to advanced therapies [1]. Nevertheless, automatic PPIMs extraction of biological relationships from biomedical literature aims to detect PPIM mentions within a (natural language) text, but no clues about the theory an experiment is proving, are specified [7,12,15].

There has been an increasing number of worldwide research groups working on text mining in the biological domain, and significant progress has been made within the *BioCreative*[1]. In particular, two challenges are relevant for PPIM mentions extraction: *BioCreative II* for task 1A (gene mention tagging), task 1B (Human Gene Normalization) and task 2 (Protein–Protein Interactions), and *BioCreative III* for task GN (Gene Normalization), IAT (Interactive Demonstration Task for Gene Indexing and Retrieval) and PPI (Protein–Protein Interactions).

Popular approaches using *BioCreative* test corpus for PPIM mentions extraction are based on linguistic pattern matching techniques [7] whereas others use a dictionary containing synonyms for PPIM names within the literature. Experimental results show a *Precision* of 10.30% (i.e., proportion of PPIM mentions that are correctly detected) and a *Recall* of 77.60% (i.e., proportion of detected PPIM mentions that correspond to existing PPIMs) where low recall can partially be due to the few PPIMs mentioned in the texts because they are not in the dictionary. Another problem is that human experts have their own writing styles for mentioning PPIMs (e.g., the PPIM "*affinity chromatography technology*" is also mentioned as "*affinity chroma*" which is not contained in the dictionary), which might be addressed by using the previous word-based criteria.

Slight increases of detection performance have been observed when using simple statistical techniques to select the best PPIM mentions [12]. These assign a weight which is proportional to the probability certain words are contained in the PPIM mention within a text. Some variations of this approach detect PPIMs referencing unknown words by calculating the probability of occurrence of a word within sample texts but including those texts that do not mention a target PPIM. Results indicate a better *Precision* of 16.60% but a lower *Recall* of 69.46%, suggesting there are no clear deterministic rules to detect PPIM mentions. To address this, machine learning has been used to automatically learn the best models to determine whether certain words refers to a PPIM.

SVM classifiers are trained using corpus containing samples of PPIM and non-PPIM mentions, surrounding words, synonyms, etc [16]. Compared to statistical techniques, the results show a significant increase of *Precision* (72.12%) and *Recall* (51.31%). However, many words are incorrectly associated with PPIM names, which may be due to that names contain too specific surrounding words that cannot be generalized to other PPIMs. Furthermore, while the approach detects the existence of a PPIM in a document, it is unable to determine the specific name. On the other hand, LR assumes that several independent variables are effective for classification. Independent variables are obtained from training texts including PPIM

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