Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Using distant supervised learning to identify protein subcellular localizations from full-text scientific articles

Wu Zheng [a], Catherine Blake [b,*]

[a] Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, USA
[b] Graduate School of Library and Information Science and Medical Information Science, Center for Informatics Research in Science and Scholarship (CIRSS), University of Illinois at Urbana-Champaign, USA

ARTICLE INFO

Article history:
Received 24 November 2014
Revised 14 July 2015
Accepted 15 July 2015
Available online 26 July 2015

Keywords:
BioNLP
Text mining
Relation extraction
Distant supervised learning
Protein subcellular localization extraction

ABSTRACT

Databases of curated biomedical knowledge, such as the protein–locations reflected in the UniProtKB database, provide an accurate and useful resource to researchers and decision makers. Our goal is to augment the manual efforts currently used to curate knowledge bases with automated approaches that leverage the increased availability of full-text scientific articles. This paper describes experiments that use distant supervised learning to identify protein subcellular localizations, which are important to understand protein function and to identify candidate drug targets. Experiments consider Swiss-Prot, the manually annotated subset of the UniProtKB protein knowledge base, and 43,000 full-text articles from the Journal of Biological Chemistry that contain just under 11.5 million sentences. The system achieves 0.81 precision and 0.49 recall at sentence level and an accuracy of 57% on held-out instances in a test set. Moreover, the approach identifies 8210 instances that are not in the UniProtKB knowledge base. Manual inspection of the 50 most likely relations showed that 41 (82%) were valid. These results have immediate benefit to researchers interested in protein function, and suggest that distant supervision should be explored to complement other manual data curation efforts.

1. Introduction

The quantity of research literature in biomedical domain is increasing at alarming rate, for example the number of abstracts in Pubmed has grown by 2 million since 2011. Although published articles capture precious information, the huge quantity of information makes it difficult for a researcher to navigate and interpret results. Text mining tools can help to alleviate this knowledge acquisition problem by extracting structured information from unstructured text. Biomedical natural language processing (BioNLP) projects used to consider only the abstracts or a limited number of full-text articles. For example, the GENIA corpus [1], one of the most widely used corpora in BioNLP contains 2000 abstracts. The focus on abstracts is in part because abstracts are more accessible and in part because it is difficult and time-consuming to annotate full-text corpora. However, previous studies have shown that abstracts provide fewer than 8% of the claims made in an articles [2,3], and that the structural, linguistic, and semantic aspects of abstracts and full-text articles markedly differ [4]. Thus, the performance of BioNLP systems that use only abstracts may not generalize to large full-text data sets.

Thanks to the development of information technology and the widespread adoption of open-access publishing, the availability of full-text scientific literature is increasing exponentially. On the other hand, realizing the gap on system performances between experimental abstracts and real-life full-text articles, BioNLP researchers gradually shift their focus from abstracts to full-text corpora. For example, Verspoor et al. developed the CRAFT corpus [5] that has 97 full-text journal articles. The last two BioNLP shared tasks have both include full-text articles in the data set [6,7]. Torii et al. [8] created a collection of 100 manually annotated full-text articles to test their rule-based information extraction system on Protein Phosphorylation. Van Landeghem et al. [9] developed an event extraction and gene normalization system using PubMed Central open access full-text collection.

In this paper, we report experimental results of automatic extraction of protein sub-cellular localization information from a large number of full-text articles using distant supervised learning. This approach leverages both full-text articles such as those that are now available in PubMed Central, and human curated knowledge bases such as the Gene Ontology [10] and UniProtKB [11]

* Corresponding author at: Graduate School of Library and Information Science and Medical Information Science, Center for Informatics Research in Science and Scholarship (CIRSS), University of Illinois at Urbana Champaign, 501 E Daniel Street, MC-493, Champaign, IL 61820-6211, USA. Tel.: +1 217 333 0115; fax: +1 217 244 3302.

E-mail addresses: wuzheng2@illinois.edu (W. Zheng), clblake@illinois.edu (C. Blake).

http://dx.doi.org/10.1016/j.jbi.2015.07.013
1532-0464/© 2015 Elsevier Inc. All rights reserved.

which provide rich, although incomplete, information about genes and proteins respectively.

The intuition is that relations stored in a knowledge base will be mentioned in many sentences in the text corpus. A BioNLP system can find these sentences, extract features indicative of the relations, and learn a classifier based on these features. The biggest advantage of this approach is that it does not require annotated text as training data. Instead, existing knowledge bases provide positive examples of the relations of interest. Since human annotation is not necessary, it is much easier to train a system on large full-text corpus. In this study, we focus extracting protein subcellular localization relations, which are important to understand protein functions and identify drug targets. The goal of this paper is to extract protein localization information from full-text scientific articles.

The rest of the paper is organized as follows. Section 2 presents related work on both protein subcellular localization extraction and the distant supervised learning approach. Section 3 describes the method in detail including the knowledge base, text corpus, and system implementation. Section 4 provides the experimental results and places those results in the contexts of other biomedical applications. Section 5 concludes the study.

2. Related work

The earliest work on protein subcellular localization extraction from text dates back to 1999 [12]. Craven and Kumlien used a Naïve Bayes classifier with bag of word (BOW) representation of text. They trained two classifiers. The first classifier was trained on a set of 2889 human annotated MEDLINE abstracts. The second classifier may be the earliest application of distant supervised learning in the biomedical domain. To be specific, they aligned the Yeast Protein Database (YPD) [13] with 633 abstract sentences describing the relation instances in YPD to get positive examples. The first classifier achieved 0.7 precision at 0.25 recall and the second classifier got 0.92 precision at 0.21 recall. The second way of training in this work is very similar to our approach, but instead of abstracts, we use a corpus of 43,072 full-text articles.

Protein sub-cellular localization is also one of the 13 biomedical relations targeted by the Genia event task [6] of the BioNLP shared task. This task differs from our study in that, first, it is framed as a supervised learning problem on limited amount of human annotated text, whereas our study is a distant supervised learning on large free text corpora. Second, the Genia task does not require the presence of both protein and location within the same sentence while our study does.

Researchers have also explored subcellular locations for genes rather than proteins. For example, the GETM system [14], was initially developed to identify gene locations from abstracts and then subsequently applied to full text articles. The system couples entity identification (gene names, anatomical terms and cell lines, from existing ontologies and tools) with a manually created set of trigger terms. The authors evaluated the system on 150 abstracts (an extension of the BioNLP '09 corpora) comprising 377 gene expressions. Of the gene expression, the authors could link 267 to anatomical locations. Their error analysis identified missing gene name as the primary source of errors (50% of false negatives).

A related but from a different field of research is protein subcellular localization predictions that use sequence information. For example, the MultiLoc system [15] uses such sequences, motifs and amino acid composition and other work has extended these sources to include protein networks [16]. The SherLoc system [17] extends earlier work by adding text to the non-text features and the EpiLoc system [18] relates most closely to the work presented here because the system uses only text-based features. Specifically EpiLoc represents text from Medline abstracts as a vector of terms and uses a support vector machine to predict the most likely location for a new protein. The authors conducted a series of experiments that compared the classification accuracy between (i) four plants and three non-plant locations, (ii) eleven locations, (iii) different weights (odds ratio, chi-squared, mutual information and information gain), and (iv) different datasets. Classification accuracy varied between species, locations, and datasets. The TargetP system also uses sequence data, but employs a neural network approach [19]. The PLOC system uses features from the amino acid, amino acid pair, and gapped amino acid pair with a support vector machine to predict 12 subcelluar locations [20]. The location accuracy for these systems was 91.5% for EpiLoc, 85.8% for TargetP [19] dataset, 78.2% for the PLOC [20] dataset and 98.7% for the MultiLoc [15] dataset. The goal of these works is different from ours in that the protein subcellular localization information the system tries to predict is not mentioned in text, whereas our goal is to extract this information from text.

With respect to distant supervision, our work is inspired by Mintz et al. [21], who used the Freebase [22] knowledge base to extract 10,000 instances of 102 relations from Wikipedia articles at a precision of 67.6%. Their evaluation assumed that any sentence containing a pair of named entities (person, organization, location) that participate in a known relation is likely to express the relation in some way. They also reported that dependency paths were particularly helpful for this task.

Riedel et al. [23] observed that the performance of distant supervision drops when the knowledge base is not tightly aligned to the text corpus. To solve this problem, they cast distant supervision as a form of multi-instance learning, assuming that in all sentences that contain a certain pair of entities which participate in a known relation, at least one sentence expresses the relation. They extracted the same relations as Mintz et al. [21] from a New York Times Corpus using Freebase as external knowledge base. They reported that their algorithm achieved 31% error reduction over the original distant supervision algorithm.

Extending the work of Riedel et al. [23], Hoffmann et al. [24] proposed an approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. They used the same text corpus and knowledge base as Riedel et al. [23] and reported gains in accuracy at both the aggregate and sentence level.

There are also applications of distant supervised learning in the biomedical domain. Ravikumar et al. [25,26] used Protein Data Bank (PDB) [27] to create training examples from which rules are learned to extract protein-residue associations. The approach adopted in this work is very similar to ours. First, they use distant supervised learning to create training and testing data. To be specific, for each entry in PDB, they get all the cited PubMed abstracts. Protein names are detected using dictionary loop-up and residues using regular expression. Sentences with protein-residue pairs that are documented in PDB are considered positive examples. In our work, we use the same approach to build our training and testing corpora. We use UniProtKB as our external knowledge source and align it with our text collection to create positive examples. Second, both Ravikumar et al. and this system rely on the dependency parse of a sentence to find target relations. After the creation of the training set, Ravikumar et al. extracted the shortest dependency paths connecting the two entities as potential rules for protein-residue association. Rule set optimization was performed to exclude rules that generate too many false positives in the training set. The optimized rules were used to detect protein-residue association in new text. In a later article, Liu et al. [28] developed an approximate matching algorithm to allow partial match of the rules with the dependency paths in new sentences. In our work,