Corpus Annotation of Biomedical Journals for a Machine Learning Approach to Automatic Relation Extraction

Andrew Laron
De La Salle University
andrew_v_laron@dlsu.edu.ph

Daniel Stanley Tan
De La Salle University
daniel.tan@dlsu.edu.ph

Nathalie Rose Lim-Cheng
De La Salle University
nats.lim@delasalle.ph

Angelyn Lao
De La Salle University
angelyn.lao@dlsu.edu.ph

Riza Batista-Navarro
University of Manchester
riza.batista@manchester.ac.uk

ABSTRACT

Predicting novel associations between biomedical entities can suggest new topics for experiments and new insights in drug design. Due to the massive amounts of relevant data available, a computational approach is well-suited for this task.

Existing studies on predicting associations between biomedical entities based on articles often use a co-occurrence-based approach, assigning weights to associations between extracted entities based on how many times entity pairs occur in different documents.

A semantic analysis-based approach may provide more detailed information to biomedical researchers. Such an approach identifies general associations between entities if the actual words of the text suggest a relation. Accounting for the words surrounding the entity terms also allows the relation to be categorized. However, it also requires a more detailed annotation schema.

This paper describes an annotation schema for automatic extraction of categorized binary relations between biomedical entities, as well as a design for a relation extraction system which makes use of this schema and has influenced its design. It also describes the software tools used to assist in the annotation process, along with issues and initial results from corpus annotation.

CCS CONCEPTS

•**Information systems** → **Information extraction;** *Chemical and biochemical retrieval;* •**Applied computing** → **Bioinformatics;** •**Computing methodologies** → *Supervised learning by classification;*

ADDITIONAL KEY WORDS AND PHRASES

relation extraction, biomedical entity extraction

1 INTRODUCTION

Biomedical researchers have conducted various studies on the relationships between human diseases [18, 27], the relationships and interactions among human genes and proteins [4], and the associations between genes and diseases [9, 23]. Identifying these associations is well-suited to a computational approach, due to the complexity of biochemical processes and the wide variety of known human diseases and disorders. Knowledge of one set of relationships can be used to produce a predicted set of other relationships.

Studies have been conducted on identifying associations between biomedical terms from article texts using co-occurrence as a measure of association. However, the co-occurrence based approach is limited: it assumes association between two entities exists as long as they appear in the sentence, when this is not always the case. Sometimes, two entities appear in a sentence that specifically states that there is no relation between them. In addition, even if it correctly identifies the associations, it provides no information on their nature.

Using natural language processing techniques, such as relation extraction, can allow researchers to identify the presence or absence of relations between entities. Some works have already been published using such techniques, including pattern matching [25], conditional random fields [3], and other approaches. Event extraction systems, which consider not only the entities in a relationship but also a *trigger word* that marks an event, have also been developed [6].

However, to identify categories of relations, an automated relation extraction system will require a more detailed model of relations than co-occurrence, as well as an annotated corpus containing examples of these relations. Having more nuanced information on relation types and strengths of associations may allow for the creation of detailed visualizations and search tools for querying this data.

We briefly discuss related works in section 2. In section 4, we discuss the relation extraction system to be used for automated annotation. In section 3, we discuss the categories and attributes of annotations and show some examples of annotated relations. In section 5, we discuss the tools and methods used for manual annotation. Section 6 describes the results and issues encountered in annotation. Lastly, we list our conclusions in section 7.

2 RELATED WORKS

2.1 Biomedical information extraction systems

Existing systems for mining biomedical relations from article text include DISEASES [20], which uses a disease-gene association system with a modified co-occurrence algorithm to assign higher weights to relations between two entities if fewer other entities occur in the same sentence.

Another system [26] extracts gene interactions from research articles using a dictionary of "interaction words". Sentences containing genes and interaction words are parsed with a dependency parser. Then, an SVM is used to determine whether sentences described a gene interaction or not, making use of a kernel function based on edit distance between test sentences and sentences known to describe gene interactions.

Philippine Computing Journal Dedicated Issue on Natural Language Processing, pages 8–15
Vol. XI No. 1 August 2016