Joint Extraction of Compound Entities and Relationships from Biomedical Literature

Cartic Ramakrishnan, Pablo N. Mendes, Rodrigo A.T.S da Gama,
Guilherme C. N. Ferreira & Amit P. Sheth
Kno.e.sis Center, Dept. of Computer Science & Engineering,
Wright State University 3640 Colonel Glenn Hwy. Dayton, Ohio
{ramakrishnan.4,mendes.2,siqueiradagama.2}@wright.edu
{denapoli.2,amit.sheth}@wright.edu

Abstract

In this paper we identify some limitations of contemporary information extraction mechanisms in the context of biomedical literature. We present an extraction mechanism that generates structured representations of textual content. Our extraction mechanism achieves this by extracting compound entities, and relationships between them, occuring in text. A detailed evaluation of the relationship and compound entities extracted is presented. Our results show over 62% average precision across 8 relationship types tested with over 82% average precision for compound entity identification[1].

1 Introduction

Contemporary search engines have harnessed the hyperlink structure of the web to provide very accurate search results for keyword searches. Information requests in the form of keywords often return relevant documents within the first page of hits. However, recent work has focussed on aggregating content across all search results, to group different documents into storylines or threads thereby implicitly connecting documents [1]. In work along similar lines, Guha et. al. [2] introduced the notion of a "Research Search" as a type of Semantic Search where users start with a search phrase which refers to an entity and this search method gathers pieces of information from multiple documents which collectively satisfy their information need. The MEMEX vision outlined by Vannevar Bush in 1945 [4] describes just such aggregation operations over text leading to insight and discovery. We believe that this largely unrealized, longstanding vision of such aggregation operations is symptomatic of a dire need for knowledge discovery operations over text. This need is most apparent when we consider text databases such as PubMed[2] where there are millions of abstracts of scientific articles which are devoid of hyperlinks. Complex interactions hidden in fragments of text are spread across several documents in scientific literature. The inability of information extraction engines to extract and interpret these complex structures therefore result in a loss of valuable knowledge. Effective utilization of these fragments to support knowledge discovery therefore requires the extraction and aggregation of knowledge. This knowledge hidden in text can be seen as collections of named relationships connecting entities. We envision a system that supports flexible expert-stipulated knowledge discovery over text. As a step towards realizing this vision, in this paper we present a mechanism for joint extraction of compound entities and relatiosnhips between them. We focus on evluating the quality of extraction results via manual evaluation to assess the effectiveness of our extraction mechanism.

2 Related Work

Supervised approaches to entity identification, or named entity recognition (NER), typically utilize training data in the form of manually labeled corpora, with tags marking entity mentions [5], [6]. Corpora such as [5] and [6] contain labeled entity mentions (e.g. lupus, autoantibodies etc. in Figure 1). Such tagged corpora are used to collect orthographical [8], contextual [9] and lexical features [10], among others. These features have been shown to perform very well in sequential labeling approaches [10] for identifying specific types of entities, like gene names,

[1]This research was supported by NSF-ITR Awards #IIS-0325464 and #0714441 titled "SemDIS: Discovering Complex Relationships in the Semantic Web."

[2]http://www.ncbi.nlm.nih.gov/pubmed/

protein names etc. [8] In these cases the types of entities sought were known and consequently a limited number of atomic observations encoded as features sufficed to identify these entities. However, a quick look at sentences in these corpora shows that token sequences marked as entities are often contained within larger logical entities that are themselves unmarked. This necessitates compound entity identification and complex relationship identification.

3 Our Approach and Contributions

In this paper:

1. We present a mechanism to jointly extract compound entities and relationships between them.
2. We manually evaluate the extracted entity-relationship-entity triples to assess the accuracy of our extraction mechanism.

The quality of the extraction of entities and relationships is of utmost importance. Our approach to extraction and aggregation in this paper is based on the key observation that entities in biomedical text are often structurally and semantically complex thereby making the relationships between them complex. These compound entities are often composed of simpler entities such as names of diseases, body parts, processes and substances. As a consequence of this we hypothesize that this percieved complexity of entities can be leveraged to identify complex relationships and interactions.

4 Extraction Algorithm

The main idea behind our extraction algorithm is to segment a given sentence into the subject ,predicate and object triples. We use rules over dependency relations to determine token sequences that together compose compound entities and relationships. Using the Stanford parser [12] we collect dependencies between tokens in each input sentence. Iterating over the dependencies, we mark words as either dominant terms (also referred to as entity/relationship "heads"), or entity/relationship modifiers. Following this step we then establish connections between heads to form triples and attach modifiers to their corresponding heads. The Stanford dependency scheme contains 48 grammatical relations organized in a hierarchy. We focus our attention mainly on the argument, conjunct, auxiliary and modifier dependency types. Evidence presented by Carroll et. al. [13] suggests that the dependency types handled by our rules are the most frequently occurring. We use the example in Figure 1 to describe our rules. The figure shows a sentence from the

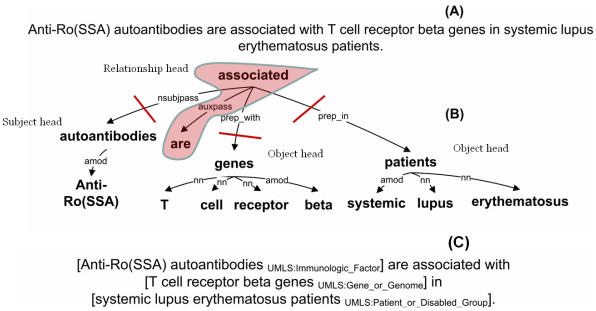

Figure 1. Sample sentence with complex relationship

GENIA[3] corpus. This sentence shows a simple case when the GENIA annotations mark compound entities correctly. Subsequent examples will deal with the case when entities identified by our method are different from those in corpora such as BioInfer and GENIA. We process dependency trees to determine cut points. The dependency types that trigger rules for this tree are shown in Figure 1. The nsubjpass results in the classification of "autoantibodies" as a compound entity head and "associated" as a predicate head. Therefore the link between "autoantibodies" and "associated" indicates that a compound entity governed by "autoantibodies" play the subject role of the predicate "associated". Similarly with auxpass, part-of-speech tests on the two words in this dependency triggers an association that the word "are" is a modifier of the relationship "associated". The dependencies prep_with and prep_in describe relational roles associated_with and associated_in, between the relationship "associated" and their dependents ("genes" and "patients"). The words genes and patients are recorded as the syntactic heads of candidate compound entities playing the object role in this sentence. Having recorded these role specific connections between relationships and their subject/object, we recursively expand the heads of candidate compound entities collecting modifiers to compose the token sequence that makes up each compound entity. Since dependency parses are not guaranteed to be acyclic we terminate the recursive expansion when we detect cycles. The recursive expansion procedure results in the entities "T cell receptor beta genes" and "systemic lupus erythematosus patients".

4.1 Rules

In order to minimize the number of rules encoded we use the hierarchy of dependencies provided by the Stanford parser. Dependency types are organized in a hierarchy

[3]This sentence is in the Genia corpus version 3.02. This sentence is the title of the abstract 90110496 in GENIA.

based on similarity in their grammatical roles. We consider a dependency d to belong to a dependency type C if d is located under C in the dependency hierarchy. This affords us the generalization capability needed to reduce the rule space. We iterate over all edges of a dependency parse tree and use the following rules to segment sentences:

1. If a dependency d(w_1, w_2) is within the dependency class SUBJECT, we mark w_2 as a head of a subject and w_1 as a head of a predicate

2. If a dependency d(w_1, w_2) is within the dependency class COMPLEMENT, we mark w_1 as a head of a predicate and w_2 as a head of an object. e.g. dobj$(w_1 = induces, w_2 = hyperplasia)$.

3. If a dependency d(w_1, w_2) is within the dependency class PREPOSITION, and w_1 is a verb, we mark w_2 as the head of an object, w_1 as a head of a predicate and combine it with the preposition (e.g. prep_with$(associated, genes)$ results in "associated with" and "genes"). If w_1 is not a verb, we combine w_1 and w_2 as a compound entity. e.g. prep_of$(w_1 = hyperplasia, w_2 = endometrium)$ results in "hyperplasia of endometrium".

Using the rules above we run relationship and compound entity extraction over text and convert the resulting triples into RDF. We generate RDF according to the scheme proposed in [11].

5 Experiments & Results

In this paper we used text from OMIM[4] as our dataset. Using the query term "renal" against OMIM, we collected the 1248 records pertaining to disease phenotypes returned by this query. Splitting these text records into sentences resulted in 71,325 sentences. We use this set for our experiments in this paper. Running our extraction algorithm on these sentences resulted in 188 Megabytes of RDF, containing approximately 150K triples. An ideal evaluation for the relationships and compound entity extraction proposed here would involve comparison with respect to existing manually annotated corpora. Our work in this paper draws a distinction between the identification of entity mentions and their relationships versus compound entities and their relationships expressed in the sentence. In order to provide a quantitative evaluation of our extraction quality we have built an evaluation tool that allows the user to perform a per-predicate evaluation. The generated RDF is loaded into a Jena[5] model. The ARQ[6] query language

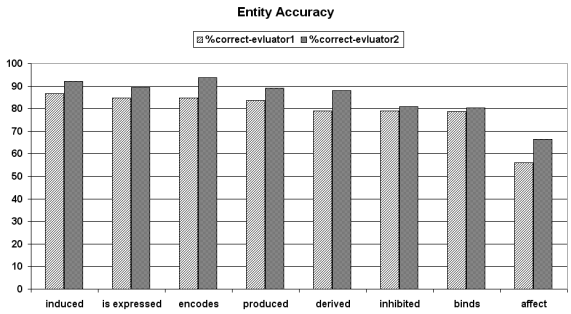

Figure 2. Compound entity % correctness across relationship types compared across evaluators

extension for SPARQL is used to formulate queries over the model and generate a list of relationship names sorted on their frequency of occurence in the RDF. From this sorted list of relationships we picked the following relationships at random from among the more frequent ones: encodes(397), is expressed(356), induced(305), produced(221), inhibited(181), derived(172), affect(166), binds(140). Our evaluation tool allows the user to iterate over the triples involving each relationship selected from the list above and juxtaposes the original sentence from which the triple was extracted, with the triple. The user is therefore able to see whether the entities and the triple involving them are indeed correct. The user rates each subject, object and the triple on binary a rating system (correct/incorrect). In our experiment each user therefore evaluates a total of 1938 instances of triples and their corresponding subject-object pairs corresponding to the 8 relationship types shown in the table above. The process of deciding whether an entity/triple is correct is based on the reader's interpretation of the sentence but follows some generic rules. A valid entity should be treated as correct if it does not need any other words from the sentence to describe it correctly and does not have any unnecessary words, which do not refer to the head of the entity. A valid triple should be treated as correct if it has the correct and full subject/object, or some word that represents it (such as prepositions). Another objective of our experiment in this paper is to see if generality or specificity of a relationship affects the accuracy with which a triple containing that relationship is extracted. While the sparsity of rules used by our system affects this accuracy in general, evidence presented by Carroll et. al. [13] suggests that our rules do cover a majority of the dependency types. It therefore seems likely that we might be able to show the effect of relationship specificity on extraction quality. Our long-term goal in studying these differences is to try and map them onto linguistic patterns that are indicative of certain

[4]http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
[5]http://jena.sourceforge.net/
[6]http://jena.sourceforge.net/ARQ/

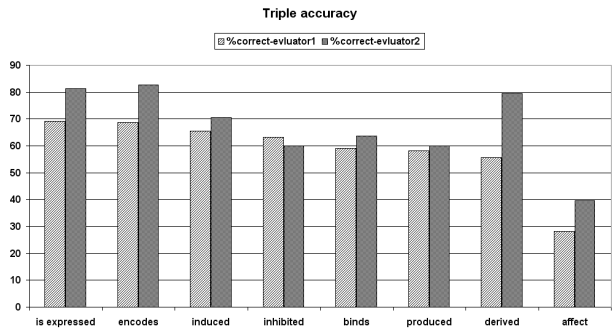

Figure 3. Triple extraction % correctness across relationship types compared across evaluators

relationships and entity types, thereby enabling the creation of type specific transducers for entity and relationship extraction. The results in Figure 3 and Figure 2 show precision comparison of triples and the entities respectively across the 8 relationship types for each evluator. A close look at the results of our experiments in Figure 3 shows that the relationships "encodes", "is expressed" and "induced" are extracted with much higher precision than "affect", "binds" and "produced". The results in Figure 2 show that the accurate identification of entities is closely tied to the relationship type. In other words, when domain-specific relationships such as "encodes", "is expressed" and "induced" occur in a sentence, our rules are able to identify entities more accurately than in the cases where general relationships such as "affect" occur. In our experiment we discovered 46,490 distinct predicates. Many of these are variants of other predicates. Normalizing these variations might allow us to get stronger patterns indicating entity prediction accuracy being affected by relationship type. However this variant normalization of predicates is not a trivial task and beyond the scope of this paper.

References

[1] Kumar, R., Mahadevan, U., and Sivakumar, D. 2004. A graph-theoretic approach to extract storylines from search results. In Proceedings of the Tenth ACM SIGKDD (Seattle, WA, USA, August 22 - 25, 2004). KDD '04.

[2] Guha, R., R. McCool, and E. Miller, Semantic search, in WWW '03 p. 700-709.

[3] Mei, Q. and Zhai, C. 2005. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In Proceedings of the Eleventh ACM SIGKDD (Chicago, Illinois, USA, August 21 - 24, 2005). KDD '05.

[4] Bush, V., As We May Think. The Atlantic Monthly, 1945. 176(1): p. 101-108.

[5] Kim, J.D., et al., GENIA corpus–semantically annotated corpus for bio-textmining. Bioinformatics, 2003. 19 Suppl 1.

[6] Pyysalo, S., et al., BioInfer: A corpus for information extraction in the biomedical domain. BMC Bioinformatics, 2007. 8(1).

[7] Alex, B., B. Haddow, and C. Grover, Recognising Nested Named Entities in Biomedical Text, in BioNLP 2007: Biological, translational, and clinical language processing. 2007: Prague.

[8] Tsai, R., et al., NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. BMC Bioinformatics 2006, 2006. 7(5).

[9] Talukdar, P., et al. A Context Pattern Induction Method for Named Entity Extraction. in CoNLL-X. 2006.

[10] McCallum, A. and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. in Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. 2003: Association for Computational Linguistics.

[11] C. Ramakrishnan, K. J. Kochut and A.P. Sheth *A Framework for Schema-Driven Relationship Discovery from Unstructured Text*, ISWC 2006: pp 583-596

[12] Klein, D. and C. Manning, Fast exact inference with a factored model for natural language parsing, in NIPS. 2003.

[13] Carrol, J., Minnen, G. and Briscoe, T. (1999) Corpus annotation for parser evaluation, Journe(s) ATALA sur les corpus annots pour la syntaxe. Paris, France.

[14] Rosario, B. and M. Hearst. Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy. in EMNLP 2001.

[15] Cartic Ramakrishnan, W. H. Milnor, M. Perry and A. P. Sheth *Discovering informative connection subgraphs in multi-relational graphs*, SIGKDD Explorations 7(2): p. 56-63 (2005).