2008

Unsupervised Discovery of Compound Entities for Relationship Extraction

Cartic Ramakrishnan
Wright State University - Main Campus

Pablo N. Mendes
Wright State University - Main Campus

Shaojun Wang
Wright State University - Main Campus, shaojun.wang@wright.edu

Amit P. Sheth
Wright State University - Main Campus, amit@sc.edu

Repository Citation

Ramakrishnan, C., Mendes, P. N., Wang, S., & Sheth, A. P. (2008). Unsupervised Discovery of Compound Entities for Relationship Extraction. *Lecture Notes in Computer Science, 5268*, 146-155.
https://corescholar.libraries.wright.edu/knoesis/49

Unsupervised Discovery of Compound Entities for Relationship Extraction

Cartic Ramakrishnan, Pablo N. Mendes, Shaojun Wang & 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,shaojun.wang,amit.sheth}@wright.edu
http://knoesis.wright.edu/

Abstract. In this paper we investigate unsupervised population of a biomedical ontology via information extraction from biomedical literature. Relationships in text seldom connect simple entities. We therefore focus on identifying compound entities rather than mentions of simple entities. We present a method based on rules over grammatical dependency structures for unsupervised segmentation of sentences into compound entities and relationships. We complement the rule-based approach with a statistical component that prunes structures with low information content, thereby reducing false positives in the prediction of compound entities, their constituents and relationships. The extraction is manually evaluated with respect to the UMLS Semantic Network by analyzing the conformance of the extracted triples with the corresponding UMLS relationship type definitions.

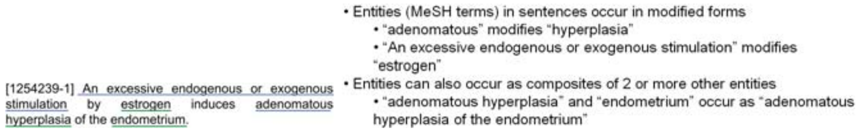
Key words: Information extraction, compound entity identification, relationship extraction, relational knowledge acquisition

1 Introduction

It is clear that there are large bodies of knowledge in textual form (e.g. PubMed [1]) that can be utilized in a variety of applications. PubMed is a service of the U.S. National Library of Medicine that includes over 17 million abstracts from life science journals that have been growing at a phenomenal rate. Consequently, the amount of Undiscovered Public Knowledge [22] is also likely to increase at a comparable rate. Motivated by this, future Semantic Web applications will seek to support semi-automated hypothesis validation directly over textual content. These operations will obviate the need for the user to sift through massive ranked lists of documents while seeking validation of their hypotheses. Instead hypotheses will be tested against knowledge aggregated from multiple texts. Affecting such aggregation of textual data will require the identification of entities, their syntactic or semantic variants and the complex combinations thereof that form concepts described in text. In addition, it is important to

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

identify relationships between these concepts so that the hypothesis connecting these entities can be validated. Relationships in text seldom manifest themselves between simple entities. Compound entities are entities that contain one or more known entities and modifiers as shown in Figure 1. The presence of these modifiers alters the semantics of compound entities necessitating the identification of their constituent entities and their types. We therefore focus on identifying compound entities rather than mentions of simple entities. To illustrate the point that entity mentions may differ from the concepts formed by their combinations, we use "hyperplasia", "endometrium" and "estrogen" as search terms resulting in a set of PubMed abstracts. One sentence from this set is shown below. In this sentence estrogen occurs in a modified form as "an excessive endogenous or exogenous stimulation by estrogen," while hyperplasia and endometrium occur in a composite form as "hyperplasia of the endometrium," further modified by the term adenomatous. This example also shows entities that are composed of non-adjacent tokens (discontinuous entities [8]). Here "an excessive endogenous" and "stimulation by estrogen" together form an entity. The example in Figure 1 shows that variants of the entities in MeSH are often found in sentences. We perceive the work in this paper as a step towards Relational Knowledge Acquisition (RKA). Our contributions can be summarized as follows:



[1254239-1] An excessive endogenous or exogenous stimulation by estrogen induces adenomatous hyperplasia of the endometrium.

- Entities (MeSH terms) in sentences occur in modified forms
  - "adenomatous" modifies "hyperplasia"
  - "An excessive endogenous or exogenous stimulation" modifies "estrogen"
- Entities can also occur as composites of 2 or more other entities
  - "adenomatous hyperplasia" and "endometrium" occur as "adenomatous hyperplasia of the endometrium"

(a) Sample Sentence (b) Modified and compound entities

Fig. 1. Sample sentence from a PubMed abstract showing compound and modified entities.

– Using a small set of rules over a dependency parse of each sentence, we segment sentences into relationships and the compound entities that form the arguments of the relationship.
– Compound entity segments are then analyzed using corpus statistics to predict their constituent entities.
– We also present a simple mechanism that uses word relations in a dependency parse to assign compound entities to a semantic class.

Existing supervised approaches to Named Entity Recognition and Relationship Extraction leverage training data that are often focused on the entity mentions. Since our approach is unsupervised it does not rely on training data and consequently does not rely on training corpora. We use GENIA [6] or BioInfer [7] sentences but do not use annotations of entities therein to guide our extraction process. Although in-depth manual curation of the extracted relationships is required, our initial results show that the extracted triples can be used for ontology population. In section 2 we present related work. In section 3 we describe our rule-based approach, leveraging the Stanford dependency scheme to segment sentences into relationships and compound entities. We present a discussion of our results in section 3 and conclude in section 4 with future work.

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 [6] and [7]. Corpora such as [6] and [7] contain labeled entity mentions (e.g. estrogen, hyperplasia etc. in Figure 1). Such tagged corpora are used to collect orthographical [9], contextual [10] and lexical features [11] among others. These features have been shown to perform very well in sequential labeling approaches [11] for identifying specific types of entities like gene names, protein names etc. [9] 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.

Recently nested and discontinuous entities [8] have received attention. The authors compare three approaches to identifying such entities through compositions of simple sequential labeling approaches viz. layering, cascading and joint labeling. They acknowledge that their approach is likely to result in prohibitively large label set when dealing with many entity types. In the biomedical domain it seems possible to find arbitrarily complex nesting of simple entities making this approach unsuitable for our purposes. For example, in Figure 1, a specific process i.e. stimulation, is the subject of the assertion. Moreover exogenous and endogenous are modifiers of this subject. These convey additional knowledge about the role of estrogen in the induction of hyperplasia. We therefore draw a distinction between the identification of entity mentions versus the meaning of the compound entities expressed in the sentence. For example, in Figure 1, “estrogen” means a biologically active substance, whereas “an excessive endogenous or exogenous stimulation by estrogen” means a biological process initiated by estrogen. It is the identification of the latter that is key in hypothesis validation.

Our approach makes it possible to perform such identification through the use of domain independent linguistic rules. We obviate supervision by using corpus based information theoretic measures to analyze the structure of compound entities. In this paper we investigate mutual information as the information theoretic measure of choice. Theoretically, other measures could be used as well.

Supervised and unsupervised approaches to relationship extraction have been attempted in the past. Machine learning approaches to the extraction of relationships between diseases and their treatments [12] have met with considerable success. Supervised approaches require the expensive human effort to create training corpora especially in the constantly evolving biomedical domain. We therefore focus our efforts on unsupervised approaches in this paper. Unsupervised approaches to relationship extraction have received considerable attention due to the training data bottleneck. Based on the interaction with a domain expert, Rinaldi et al. [14] identify a set of relations along with their morphological variants (bind, regulate, signal etc.) that are of particular interest in the biology domain. Using the dependency parse of GENIA sentences they developed a number of axioms over the dependency patterns that capture the relations that are of interest in this domain. Axiom formulation was however a manual process involving a domain expert. Other approaches have relied on hand-coded domain specific rules that encode extraction patterns used to extract molecular pathways [15] and protein interactions [16]. Ciaramita et al. [17] use the entity annotations of the GENIA [7] corpus to learn semantic relationships. In this work the authors extracted patterns indicating relationships from parse trees. The patterns themselves do not encode domain information. However they do assume prior knowledge of the entities in sentences (manual annotation from GENIA). This work provides the main motivation for our work. Because we do not assume that entity annotations are given a priori, our problem is significantly harder.

Our method is based on rules over dependency parse trees. These rules are thus agnostic to domain knowledge and our method does not require prior knowledge of entities. Thus we can apply this method to any text without having to reengineer the rules. Similar domain agnostic rules have been deployed to extract compound entities and relationships from constituency parses by Ramakrishnan et.al. [1]. The work in this paper is an extension of their work.

3 Our Approach

The main idea behind our approach is to segment dependency trees to facilitate further extraction of (subject, predicate, object) triples. We use rules over dependency relations to determine token sequences that together compose compound entities. In doing so we aim to identify and connect the appropriate entities with relationships. Using the Stanford parser [18] 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. [19] suggests that the dependency types handled by our rules are the most frequently occurring.

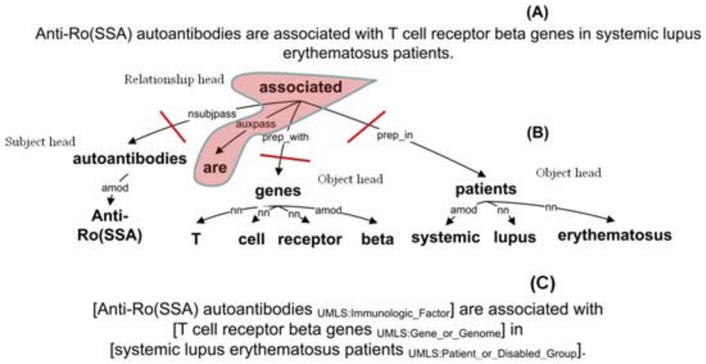

(A)
Anti-Ro(SSA) autoantibodies are associated with T cell receptor beta genes in systemic lupus erythematosus patients.
Relationship head
associated
nsubjpass
auxpass
prep_with
prep_in
(B)
Subject head
autoantibodies
are
genes
Object head
patients
Object head
amod
Anti-Ro(SSA)
nn
nn
nn
amod
T
cell
receptor
beta
amod
nn
nn
systemic
lupus
erythematosus
(C)
[Anti-Ro(SSA) autoantibodies UMLS:Immunologic_Factor] are associated with
[T cell receptor beta genes UMLS:Gene_or_Genome] in
[systemic lupus erythematosus patients UMLS:Patient_or_Disabled_Group].

Fig. 2. (A) Sample sentence (B) dependency parse tree (C) Sentence Annotated with types of compound entities discovered.

3.1 Algorithm

We use the example in Figure 2 to describe our rules. The figure shows a sentence from the GENIA[2]. 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. Consider the parse tree in Figure 2. The dependency types that trigger rules for this tree are shown in Figure 2. 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". The information recorded in this way is used by the second phase of our algorithm. In this phase the words in a compound entity are used to assign a semantic type to the compound entity. Work addressing the semantics

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

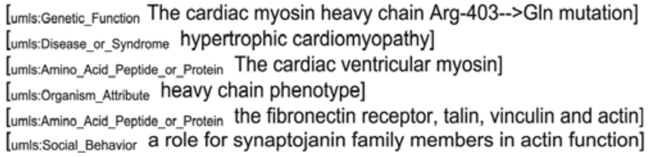
of noun compounds [20] has aimed at inferring semantic types for two word biomedical noun compounds using the MeSH hierarchy. Typing arbitrarily large noun compounds presents a significant research challenge [8]. We use the type of the compound entity head as an indicator of the possible type of the entity. We match the heads of compund entities with single-word MeSH terms. Using UMLS class that this entity belongs to, we assign that class to the compound entity. This is a simple approach to get good initial guesses of entity types. However, this may not yield correct results in all cases and further investigation extending the work in [20] is warranted. A recent approach to unsupervised extraction described in [17] relied on a sentence simplification strategy where entities (multi-word entities) were replaced with their semantic types. This resulted in a simplified parse tree and allowed for fewer rules to guide the extraction process. Our method is similar to this approach applied in the context on un-annotated corpora using dependency trees. The Stanford parser's dependency hierarchy allows for a more principled approach to reducing the number of rules.

3.2 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 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 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$(w_1 = \textit{induces}, w_2 = \textit{hyperplasia})$.
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$(\textit{associated}, \textit{genes})$ 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$(w_1 = \textit{hyperplasia}, w_2 = \textit{endometrium})$ results in "hyperplasia of endometrium".

Our initial experiment testing the utility of our rules was run on the BioInfer corpus [7]. This corpus contains 1100 sentences that have been either manually or automatically annotated. These annotations mark nested and discontinuous entities as well as relationships. Our algorithm produced 5614 entity guesses from these 1100 sentences without using the existing annotation information. The figures below show compound entities discovered by our system and segmentations of our example sentences into relationships and compound entities. The results above are obtained by using the 4 rules described in our



[umls:Genetic_Function The cardiac myosin heavy chain Arg-403-->Gln mutation]
[umls:Disease_or_Syndrome hypertrophic cardiomyopathy]
[umls:Amino_Acid_Peptide_or_Protein The cardiac ventricular myosin]
[umls:Organism_Attribute heavy chain phenotype]
[umls:Amino_Acid_Peptide_or_Protein the fibronectin receptor, talin, vinculin and actin]
[umls:Social_Behavior a role for synaptojanin family members in actin function]

Fig. 3. Compound entities identified (Note: prefix subscripted text indicates UMLS type if applicable)

algorithm. A close inspection of the entities in Figure 3 shows that all entities except the last are correct. The last entity listed in Figure 3 seems to be mistakenly tagged as an instance of UMLS class umls:Social_Behavior. This is because we use the head of the compound entity as a simple mechanism to assign a type to compound entities. Using the word "role" as the head, this entity's type is assigned. Future work will investigate this in further detail, considering issues pertaining to the semantics of noun compounds [20].

A significant proportion of the compound entities that we found were other compound entities put together using punctuations e.g. "the simultaneous quantification of myosin heavy chain, myosin light chain, phosphorylatable myosin light chain". This entity does indeed form the subject of an assertion in a BioInfer sentence[3]. However, the correct interpretation of this entity is as follows: all three types of myosin are modified by the words "simultaneous quantification". As per our previous observations, directly comparing the compound entities predicted using only the rules will show poor performance. This is due to a preponderance of entity predictions like the one discussed above. To address this issue we developed an entity prediction strategy that leverages corpus statistics to predict constituents of compound entities. One objective here is to reduce false positives in compound entity identification and the second objective is to identify the component of a compound entity and the ways in which they are combined.

3.3 Predicting constituents of Compound entities via corpus statistics

Subsequences of tokens belonging to a predicted compound entity, which co-occur across a large corpus, dependent on each other in the same manner, are likely to themselves form sub-entities. Here we do not use the typical definition of co-occurrence (i.e. adjacent terms). Instead, we use a role specific definition of co-occurrence which treats two terms as co-occurring if they are connected by a dependency of any type in a dependency parse tree. Furthermore, for this (or any form) of co-occurrence statistic to be effective, a large corpus is required.

In order to compute corpus-wide statistics we "expanded" the BioInfer corpus by increasing the number of sentences pertaining to each entity in BioInfer.

[3] See sentence with id 1610 in the BioInfer corpus @ http://mars.cs.utu.fi/BioInfer/

Using the known entities in BioInfer as seed queries, we queried the PubMed database obtaining 100 abstracts corresponding to each entity. This resulted in a set of approximately 77,000 abstracts. Splitting these abstracts into sentences yielded approximately 850,000 sentences. Using the Stanford dependency parser we parsed these 850,000 sentences. We then built a Lucene index which indexes each dependency of the form rel(gov,dep)using the name of the dependency (i.e. rel) the governor term and the dependent term (i.e. gov and dep respectively) as fields.

Mutual information [21] was introduced as a measure for discovering interesting word collocations. Intuitively, mutual information measures the information that two random variables share: it measures how much knowing one of these variables reduces our uncertainty about the other. If two variables are independent, the mutual information is zero. Mutual information increases when two words occur together very often. Consider a pair of words w_i & w_j. The pointwise mutual information between w_i & w_j, I(w_i, w_j) is computed as follows:

$$I(w_i, w_j) = p(w_i, w_j) \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \quad where \;\; p(w_i, w_j) = p(w_i)p(w_j|w_i) \tag{1}$$

The maximum likelihood estimate for p(w_i) or p(w_j) is the ratio of the frequency of occurrence of the corresponding word with the total number of words in the corpus, and the maximum likelihood estimate for the conditional probability p$(w_j|w_i)$ is the ratio of the frequency of the co-occurrence of w_i & w_j with the frequency of w_i i.e. $p(w_j|w_i) = \frac{count(w_i,w_j)}{count(w_i)}$ This definition is based on co-occurrence of words in a sentence. In our case, however the definitions of the co-occurrence counts are based on the number of dependencies connecting the two words. The idea being that, non-adjacent tokens can, in some cases, be combined to form entities. Therefore:

$$p_d(w_j|w_i) = \frac{count_d(w_i = dep \wedge w_j = gov) + count_d(w_j = dep \wedge w_i = gov)}{count_d(w_i = dep \vee w_i = gov)} \tag{2}$$

and $p_d(w_i) = \frac{count(w_i)}{N}$where N is the total number of dependencies across the entire corpus. $count_d(w_j = dep \wedge w_i = gov)$ represents the number of dependencies that have w_i & w_j as governor AND dependent respectively while $count_d(w_i = dep \vee w_i = gov)$ represents the number of dependencies in which w_i is either the governor OR the dependent.

3.4 Preliminary Results

Using this dependency-based mutual information as a guide we predict token subsequences of compound entities that are most likely to form entities themselves. Entities predicted by our algorithm are shown in Figure 3 and Figure 2. Using these compound entities as a starting point our sub-entity prediction mechanism groups tokens to form sub-entities. This results in a segmentation of compound entities into its constituents. Some results of this process are shown below.

Compound Entity	Constituent Entities Predicted
Cdc42-induced nucleation of actin filaments	nucleation of actin filaments, actin filaments
affinity of yeast profilin for rabbit actin	affinity of yeast profiling, affinity of yeast profilin for rabbit actin
main inhibitory action of p27, with the cyclin E/cyclin-dependent kinase 2 (Cdk2)	main inhibitory action, main inhibitory action of p27, with the cyclin E/cyclin-dependent kinase 2 (Cdk2)
tumor necrosis factor receptor	tumor necrosis, tumor necrosis factor receptor
actin-binding proteins of low molecular weight	actin-binding proteins, actin-binding proteins of low molecular weight
Inactivation of the Rb pathway cell lung carcinoma	cell lung carcinoma, Inactivation of the Rb pathway cell lung carcinoma
Three components of Drosophila adherens junctions	components of Drosophila adherens junctions, Drosophila adherens junctions, adherens junctions

These entities were predicted using the 1100 sentences that are in BioInfer. Corpus statistics gathered from 850,000 sentences were used to obtain entity predictions based on mutual information as discussed. In addition to entities our system predicts possible relationship triples that might hold between the entities in a sentence. The table below shows some of these relationships[4].

Relationship	Sentence Segmentation
increased	A pre-treatment of cells with SGE from partially fed ticks in amounts salivary glands → increased → the level of both viral nucleocapsid *N* protein phosphoprotein *P* in a dose-dependent manner
inhibits	alpha-catenin → inhibits → beta-catenin signaling
inhibits	MgCl2 → inhibits → these effects of profilin, most likely
causes	The cardiac myosin heavy chain Arg-403 Gln mutation → causes → hypertrophic cardiomyopathy
causes	Moreover, addition of profilin to steady-state actin filaments → causes → slow depolymerization
causes	(11-22 microM) into infected PtK2 cells → causes → a marked slowing of actin tail elongation and bacterial migration
binds	the cytoplasmic domain of E-cadherin → binds → either beta-catenin or plakoglobin
binds	a constituent → binds → RBC alpha-spectrin antibody plus the presence of significant quantities of actin

The table above clearly show the benefit of our approach. Relatively large entities with several possible sub-entities are identified by our system. Evaluation of compound entities and relationships extracted in this paper requires human subject evaluation.

4 Conclusions and Future Work

In this paper we presented a method based on rules over grammatical dependency structures for unsupervised segmentation of sentences into compound entities and relationships. This work draws a distinction between the identification of entity mentions versus the meaning of the compound entities expressed in the sentence. This makes it difficult to evaluate our results against corpora that have narrow objectives and scope. We therefore present qualitative evaluation showing the utility of results. We have also used the extracted entities and relationships in a few applications to validate the usefulness of the extraction process. The reader is invited to try out our semantic browser at this paper's resources page: http://knoesis.wright.edu/research/semweb/projects/textMining/ekaw2008.

References

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

[4] The complete set of extracted entities & relationships is available at `http://knoesis.wright.edu/research/semweb/projects/textMining/ekaw2008`.

2. 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).
3. Guha, R., R. McCool, and E. Miller, Semantic search, in WWW '03 p. 700-709.
4. Soon, W., H. Ng, and Daniel, A machine learning approach to coreference resolution of noun phrases. COLING, 2001. 27(4): p. 521-544.
5. Lappin, S. and H. Leass, An algorithm for pronominal anaphora resolution. Comput. Linguist., 1994. 20(4): p. 535-561.
6. Kim, J.D., et al., GENIA corpus–semantically annotated corpus for bio-textmining. Bioinformatics, 2003. 19 Suppl 1.
7. Pyysalo, S., et al., BioInfer: A corpus for information extraction in the biomedical domain. BMC Bioinformatics, 2007. 8(1).
8. 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.
9. 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).
10. Talukdar, P., et al. A Context Pattern Induction Method for Named Entity Extraction. in CoNLL-X. 2006.
11. McCallum, A. and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. in NLL at HLT-NAACL 2003. 2003: ACL.
12. Barbara, R. and A.H. Marti, Classifying semantic relations in bioscience texts, ACL 2004, Association for Computational Linguistics: Barcelona, Spain.
13. Mark, C. and K. Johan, Constructing Biological Knowledge Bases by Extracting Information from Text Sources, in ISMB 1999, AAAI Press.
14. Rinaldi, F., et al., Mining relations in the GENIA corpus, in Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics, held in conjunction with ECML/PKDD 2004.
15. Friedman, C., et al., GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics, 2001. 17 Suppl 1: p. 1367-4803.
16. Saric, J., et al., Extraction of regulatory gene/protein networks from Medline. Bioinformatics, 2005.
17. Ciaramita, M., et al., Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology, in 19th IJCAI 2005.
18. Klein, D. and C. Manning, Fast exact inference with a factored model for natural language parsing, in NIPS. 2003.
19. 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.
20. Rosario, B. and M. Hearst. Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy. in EMNLP 2001.
21. Church, K.W., et al., Using statistics in lexical analysis, in Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. 1991, Lawrence Erlbaum Associates.
22. D.R. Swanson *Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge.*, Perspectives in Biology and Medicine, 1986. 30(1): p. 7-18.