IntEx: A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text

Syed Toufeeq Ahmed, Deepthi Chidambaram, Hasan Davulcu*, Chitta Baral.
Department of Computer Science and Engineering,
Arizona State University,
Tempe, AZ 85287.
{toufeeq, deepthi, hdavulcu, chitta}@asu.edu

Abstract

In this paper, we present a fully automated extraction system, named IntEx, to identify gene and protein interactions in biomedical text. Our approach is based on first splitting complex sentences into simple clausal structures made up of syntactic roles. Then, tagging biological entities with the help of biomedical and linguistic ontologies. Finally, extracting complete interactions by analyzing the matching contents of syntactic roles and their linguistically significant combinations. Our extraction system handles complex sentences and extracts multiple and nested interactions specified in a sentence. Experimental evaluations with two other state of the art extraction systems indicate that the IntEx system achieves better performance without the labor intensive pattern engineering requirement.

1 Introduction

Genomic research in the last decade has resulted in the production of a large amount of data in the form of micro-array experiments, sequence information and publications discussing the discoveries. The data generated by these experiments is highly connected; the results from sequence analysis and micro-arrays depend on functional information and signal transduction pathways cited in peer-reviewed publications for evidence. Though scientists in the field are aided by many online databases of biochemical interactions, currently a majority of these are curated labor intensively by domain experts. Information extraction from text has therefore been pursued actively as an attempt to extract knowledge from published material and to speed up the curation process significantly.

In the biomedical context, the first step towards information extraction is to recognize the names of proteins (Fukuda, Tsunoda et al. 1998), genes, drugs and other molecules. The next step is to recognize interaction events between such entities (Blaschke, Andrade et al. 1999; Blaschke, Andrade et al. 1999; Hunter 2000; Thomas, Milward et al. 2000; Thomas, Rajah et al. 2000; Ono, Hishigaki et al. 2001; Hahn and Romacker 2002) and then to finally recognize the relationship between interaction events. However, several issues make extracting such interactions and relationships difficult since (Seymore, McCallum et al.1999) (i) the task involves free text – hence there are many ways of stating the same fact (ii) the genre of text is not grammatically simple (iii) the text includes a lot of technical terminology unfamiliar to existing natural language processing systems (iv) information may need to be combined across several sentences, and (v) there are many sentences from which nothing should be extracted.

In this paper, we present a fully automated extraction approach to identify gene and protein interact-

[*] To whom correspondence should be addressed

Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, pages 54–61, Detroit, June 2005. ©2005 Association for Computational Linguistics

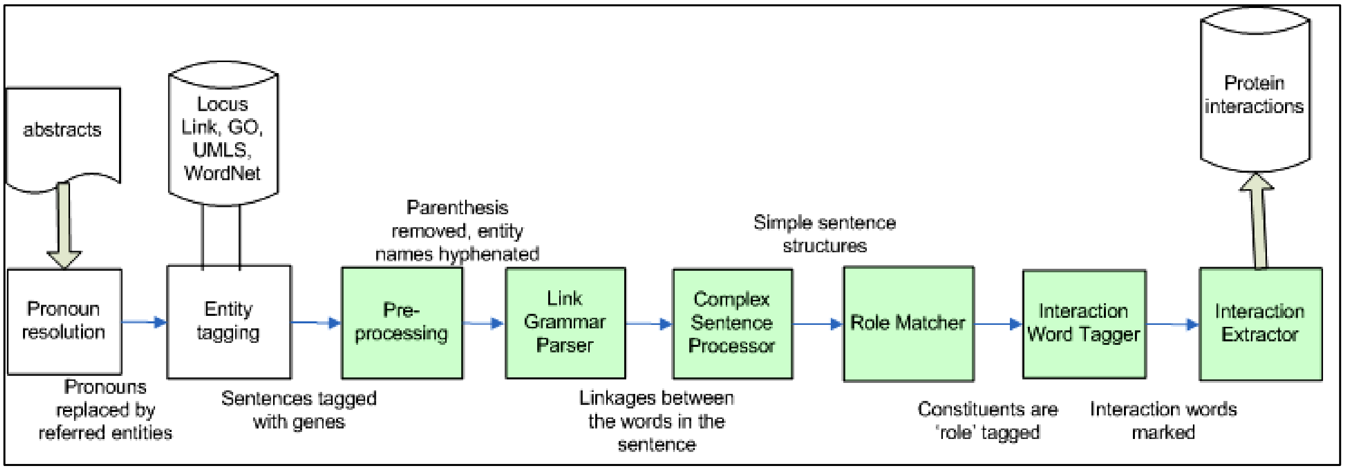

abstracts
Locus Link, GO, UMLS, WordNet
Protein interactions
Parenthesis removed, entity names hyphenated
Simple sentence structures
Pronoun resolution
Entity tagging
Pre-processing
Link Grammar Parser
Complex Sentence Processor
Role Matcher
Interaction Word Tagger
Interaction Extractor
Pronouns replaced by referred entities
Sentences tagged with genes
Linkages between the words in the sentence
Constituents are 'role' tagged
Interaction words marked

Figure 1: System Architecture

tions in natural language text with the help of biomedical and linguistic ontologies. Our approach works in three main stages:

1. **Complex Sentence Processor (CSP):** First, is splitting complex sentences into simple clausal structures made of up syntactic roles.
2. **Tagging:** Then, tagging biological entities with the help of biomedical and linguistic ontologies.
3. **Interaction Extractor:** Finally, extracting complete interactions by analyzing the matching contents of syntactic roles and their linguistically significant combinations.

The novel aspects of our system are its ability to handle complex sentence structures using the Complex Sentence Processor (CSP) and to extract multiple and nested interactions specified in a sentence using the Interaction Extractor without the labor intensive pattern engineering requirement. Our approach is based on identification of syntactic roles, such as subject, objects, verb and modifiers, by using the word dependencies. We have used a dependency based English grammar parser, the Link Grammar (Sleator and Temperley 1993), to identify the roles. Syntactic roles are utilized to transform complex sentences into their multiple clauses each containing a single event. This clausal structure enables us to engineer an automated algorithm for the extraction of events thus overcoming the burden of labor intensive pattern engineering for complex and compound sentences. Pronoun resolution module assists Interaction Extractor in identifying interactions spread across multiple sentences using pronominal references. We performed comparative experimental evaluations with two state of the art systems. Our experimental results show that the IntEx system presented here achieves better performance without the labor intensive rule engineering step which is required for these state of the art systems.

The rest of the paper is organized as follows. In Section 2 we survey the related work. In Section 3 we present an architectural overview of the IntEx system. Sections 4 and 5 explain and illustrate the individual modules of the IntEx system. A detailed evaluation of our system with the BioRAT (Corney, Buxton et al. 2004) and GeneWays (Rzhetsky, Iossifov et al. 2004) is presented in Section 6. Section 7 concludes the paper.

2 Related Work

Information extraction is the extraction of salient facts about pre-specified types of events, entities (Bunescu, Ge et al. 2003) or relationships from free text. Information extraction from free-text utilizes shallow-parsing techniques (Daelemans, Buchholz et al. 1999), Parts-of-Speech tagging(Brill 1992), noun and verb phrase chunking (Mikheev and Finch 1997), verb subject and object relationships (Daelemans, Buchholz et al. 1999), and learned (Califf and Mooney 1998; Craven and Kumlein 1999; Seymore, McCallum et al. 1999) or hand-build patterns to automate the creation of specialized databases.

Manual pattern engineering approaches employ shallow parsing with patterns to extract the interactions. In the (Ono, Hishigaki et al. 2001) system, sentences are first tagged using a dictionary based protein name identifier and then processed by a module which extracts interactions directly from complex and compound sentences using regular expressions based on part of speech tags.

The SUISEKI system of Blaschke (Blaschke, Andrade et al. 1999) also uses regular expressions, with probabilities that reflect the experimental accuracy of each pattern to extract interactions into predefined frame structures.

GENIES (Friedman, Kra et al. 2001) utilizes a grammar based NLP engine for information extraction. Recently, it has been extended as GeneWays (Rzhetsky, Iossifov et al. 2004), which also provides a Web interface that allows users to search and submit papers of interest for analysis. The BioRAT system (Corney, Buxton et al. 2004) uses manually engineered templates that combine lexical and semantic information to identify protein interactions. The GeneScene system(Leroy, Chen et al. 2003) extracts interactions using frequent preposition-based templates.

Grammar engineering approaches, on the other hand use manually generated specialized grammar rules (Rinaldi, Schneider et al. 2004) that perform a deep parse of the sentences. Temkin (Temkin and Gilder 2003) addresses the problem of extracting protein interactions by using an extendable but manually built Context Free Grammar (CFG) that is designed specifically for parsing biological text. The PathwayAssist system uses an NLP system, MedScan (Novichkova, Egorov et al. 2003), for the biomedical domain that tags the entities in text and produces a semantic tree. Slot filler type rules are engineered based on the semantic tree representation to extract relationships from text. Recently, extraction systems have also used link grammar (Grinberg, Lafferty et al. 1995) to identify interactions between proteins (Ding, Berleant et al. 2003). Their approach relies on various linkage paths between named entities such as gene and protein names. Such manual pattern engineering approaches for information extraction are very hard to scale up to large document collections since they require labor-intensive and skill-dependent pattern engineering.

Machine learning approaches have also been used to learn extraction rules from user tagged training data. These approaches represent the rules learnt in various formats such as decision trees (Chiang, Yu et al. 2004) or grammar rules (Phuong, Lee et al. 2003). Craven et al (Craven and Kumlien 1999) explored an automatic rule-learning approach that uses a combination of FOIL (Quinlan 1990) and Naïve Bayes Classifier to learn extraction rules.

3 System Architecture

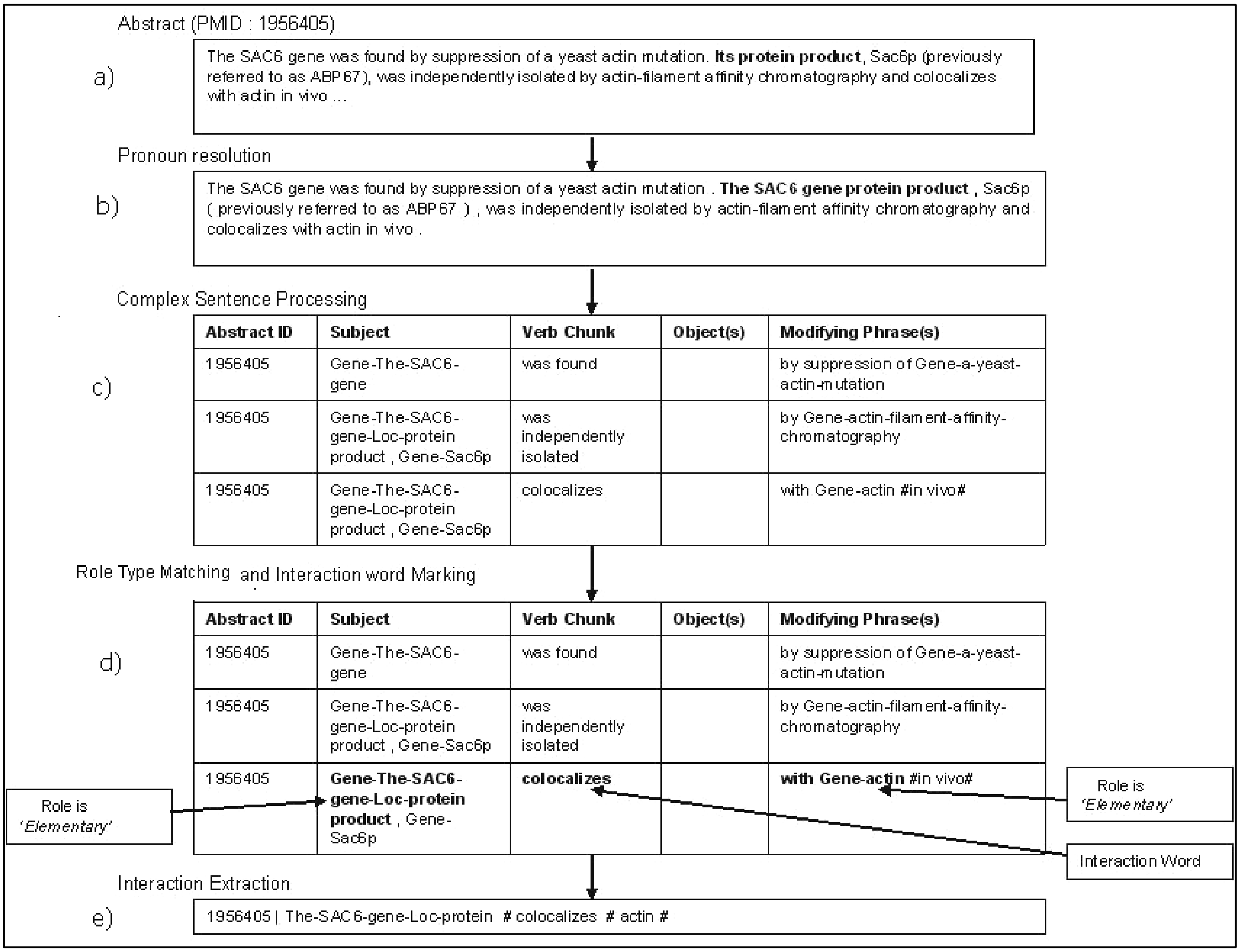
The sentences in English are classified as either simple, complex, compound or complex-compound based on the number and types of clauses present in them. Our extraction system resolves the complex, compound and complex-compound sentence structures (collectively referred to as complex sentence structures in this document) into simple sentence clauses which contain a subject and a predicate. These simple sentence clauses are then processed to obtain the interactions between proteins. The architecture of the IntEx system is shown in Figure 1, and the following Sections 4 and 5 explain the workings of its modules.

4 Complex Sentence Processing

4.1 Pronoun Resolution

Interactions are often specified through pronominal references to entities in the discourse, or through co references where, a number of phrases are used to refer to the same entity. Hence, a complete approach to extracting information from text should also take into account the resolution of these references. References to entities are generally categorized as co-references or anaphora and has been investigated using various approaches (Castaño, Zhang et al. 2002). IntEx anaphora resolution subsystem currently focuses on third person pronouns and reflexives since the first and second person pronouns are frequently used to refer to the authors of the papers.

Our pronoun resolution module uses a heuristic approach to identify the noun phrases referred by the pronouns in a sentence. The heuristic is based on the number of the pronoun (singular or plural) and the proximity of the noun phrase. The first noun phrase that matches the number of the pronoun is considered as the referred phrase.



The SAC6 gene was found by suppression of a yeast actin mutation. **Its protein product**, Sac6p (previously referred to as ABP67), was independently isolated by actin-filament affinity chromatography and colocalizes with actin in vivo ...

The SAC6 gene was found by suppression of a yeast actin mutation . **The SAC6 gene protein product** , Sac6p (previously referred to as ABP67) , was independently isolated by actin-filament affinity chromatography and colocalizes with actin in vivo .

Abstract ID	Subject	Verb Chunk	Object(s)	Modifying Phrase(s)
1956405	Gene-The-SAC6-gene	was found		by suppression of Gene-a-yeast-actin-mutation
1956405	Gene-The-SAC6-gene-Loc-protein product , Gene-Sac6p	was independently isolated		by Gene-actin-filament-affinity-chromatography
1956405	Gene-The-SAC6-gene-Loc-protein product , Gene-Sac6p	colocalizes		with Gene-actin #in vivo#

Abstract ID	Subject	Verb Chunk	Object(s)	Modifying Phrase(s)
1956405	Gene-The-SAC6-gene	was found		by suppression of Gene-a-yeast-actin-mutation
1956405	Gene-The-SAC6-gene-Loc-protein product , Gene-Sac6p	was independently isolated		by Gene-actin-filament-affinity-chromatography
1956405	**Gene-The-SAC6-gene-Loc-protein product** , Gene-Sac6p	**colocalizes**		**with Gene-actin** #in vivo#

1956405 | The-SAC6-gene-Loc-protein # colocalizes # actin #

Fig. 3 Example - a) A Sentence from an abstract (PMID: 1956405). b) Pronoun 'it's' is resolved with 'The SAC6 gene'. c) Each row represents a simple sentence, d) for each constituent, role type is resolved and interaction words are tagged, e) Protein-Protein interaction is extracted.

4.2 Entity Tagger

The entity tagging module marks the names of genes, and proteins in text. The process of tagging is a combination of dictionary look up and heuristics. Regular expressions are also used to mark the names that do not have a match in the dictionaries. The protein name dictionaries for the entity tagger are derived from various biological sources such as UMLS[1], Gene Ontology[2] and Locuslink[3] database

4.3 Preprocessor

The tagged sentences need to be pre-processed to replace syntactic constructs, such as parenthesized nouns and domain specific terminology that cause the Link Grammar Parser to produce an incorrect output. This problem is overcome by replacing such elements with alternative formats that is recognizable by the parser.

4.4 Link Grammar and the Link grammar parser

Link grammar (LG) introduced by Sleator and Temperley (Sleator and Temperley 1991) is a dependency based grammatical system. The basic idea of link grammar is to connect pairs of words

[1] http://www.nlm.nih.gov/research/umls/
[2] http://www.geneontology.org/
[3] http://www.ncbi.nlm.nih.gov/LocusLink/

in a sentence with various syntactically significant links. The LG consists of set of words, each of which has various alternative linking requirements.

A linking requirement can be seen as a block with connectors above each word. A connector is satisfied by matching it with compatible connector. Fig.2 below shows how linking requirements can be satisfied to produce a parse for the example sentence "*The dog chased a cat*".

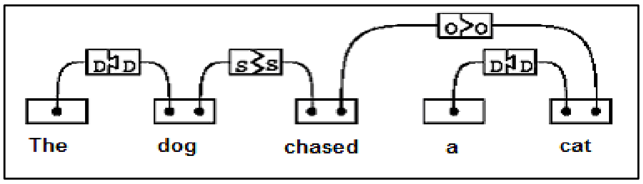


Figure 2: Link grammar representation of a sentence

Even though LG has no explicit notion of constituents or categories (Sleator and Temperley 1993), they emerge as contiguous connected sequence of words attached to the rest of sentence by a particular types of links, as in the above example where *'the dog'* and *'a cat'* are connected to the main verb via 'S' and 'O' links respectively. Our algorithms utilize this property of LG where certain link types allow us to extract the constituents of sentences irrespective of the tense. The LG parser's ability to detect multiple verbs and their constituent linkage in complex sentences makes it particularly well suited for our approach during resolving of complex sentences into their multiple clauses. The LG parsers' dictionary can also be easily enhanced to produce better parses for biomedical text (Szolovits 2003).

4.5 Complex Sentence Processor Algorithm

The complex sentence processor (CSP) component splitsthe complex sentences into a collection of simple sentence clauses which contain a subject and a predicate. The CSP follows a verb-based approach to extract the simple clauses. A sentence is identified to be complex it contains more than one verb. A simple sentence is identified to be one with a subject, a verb, objects and their modifying phrases. The example in Figure 3 illustrates the major steps involved during complex sentence processing. The following schema is used as the format to represent simple clauses:

Subject | Verb | Object | Modifying phrase to the verb

5 Interaction Extraction

Interaction Extractor (IE) extracts interactions from simple sentence clauses produced by the complex sentence processor. The highly technical terminology and the complex grammatical constructs that are present in the biomedical abstracts make the extraction task difficult, Even a simple sentence with a single verb can contain multiple and/or nested interactions. That's why our IE system is based on a deep parse tree structure presented by the LG and it considers a thorough case based analysis of contents of various syntactic roles of the sentences like their subjects (S), verbs (V), objects (O) and modifying phrases (M) as well as their linguistically significant and meaningful combinations like *S-V-O, S-O, S-V-M or S-M,* illustrated in Figure 4, for finding and extracting protein-protein interactions.

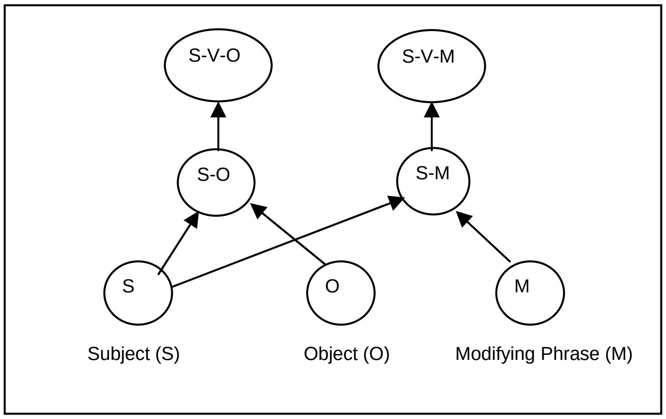


Figure 4: Interaction Extraction: Composition and analysis of various syntactic roles.

5.1 Role Type Matcher

For each syntactic constituent of the sentence, the role type matcher identifies the type of each role as either 'Elementary', 'Partial' or 'Complete' based on its matching content, as presented in Table 1.

Table 1: Role Type Matcher

Role Type	Description
Elementary	If the role contains a Protein name or an interaction word.
Partial	If the role has a Protein name and an interaction word.
Complete	If the role has at least two Protein names and an interaction word.

5.2 Interaction Word Tagger

The words that match a biologically significant action between two gene/protein names are labeled as ‘interaction words’. Our gazetteer for interaction words is derived from UMLS and WordNet[4]. Porter Stemmer (Porter 1997) was also used for stemming such words before matching.

5.3 Interaction Extractor (IE)

IntEx interaction extractor works as follows. The input to IE is the preprocessed and typed simple clause structures. The IE algorithm progresses bottom up, starting from each syntactic role S, V or M, and expanding them using the lattice provided in Figure 4 until all ‘Complete’ singleton or composite role types are obtained.

Consider the example shown in Figure 3, for the third sentence, the boundaries of the subject and the modifying phrase are identified and both are role typed as ‘Elementary’ using Table 1. Since the main verb is tagged as an interaction word, IE uses the S-V-M composite role from Figure 4 to find and extract the following complete interaction:

{‘The SAC 6 gene Protein’, ‘colocalizes’, ‘actin’}.

‘Complete’ roles also need to be analyzed in order to determine their voice as ‘active’ or ‘passive’. Since there are only a small number of preposition combinations, such as *of-by, from-to etc.*, that occur frequently within the clauses, they can be used to distinguish the agent and the theme of the interactions.

For example, in the sentence “The kinase phosphorylation of pRb by c-Abl in the gland could inhibit ku70”, the subject role is “The kinase phosphorylation of pRb by c-Abl in the gland”. Since the subject has at least two protein names and an interaction word it is ‘complete’. By using the *‘of-by’* pattern (*…<Interaction-Word (action)>... **of** ...<theme>…**by** ...<agent>…*) the IE is able to extract the correct interaction {c-Abl, phosphorylation, pRb} from the subject role alone.

6 Evaluation & discussion

We have evaluated the performance of our system with two state of the art systems - BioRAT (Corney, Buxton et al. 2004) and GeneWays (Rzhetsky, Iossifov et al. 2004).

Blaschke and Valencia (Valencia 2001) recommend DIP (Xenarios, Rice et al. 2000) dataset as a benchmark for evaluating biomedical Information Extraction systems. The first evaluation for IntEx system was performed on the same dataset [5] that was used for the BioRAT evaluation. For BioRAT evaluation, authors identified 389 interactions from the DIP database such that both proteins participating in the interaction had SwissProt entries. These interactions correspond to 229 abstracts from the PubMed. The BioRAT system was evaluated using these 229 abstracts. The interactions extracted by the system were then manually examined by a domain expert for precision and recall. Precision is a measure of correctness of the system, and is calculated as the ratio of true positives to the sum of true positives and false positives. The sensitivity of the system is given by the recall measure, calculated as the ratio of true positives to the sum of true positives and false negatives.

Table 2: Recall comparison of IntEx and BioRAT from 229 abstracts when compared with DIP database.

Recall	IntEx		BioRAT	
Results	Cases	Percent (%)	Cases	Percent(%)
Match	142	**26.94**	79	**20.31**
No Match	385	73.06	310	79.67
Totals	527	100.00	389	100.00

We have also limited our protein name dictionary to the SwissProt entries. Tables 2 and 3 present the evaluation results as compared with the BioRAT system. A detailed analysis of the sources of all types of errors is shown in Figure 6.

[4] http://www.cogsci.princeton.edu/~wn/

[5] Dataset was obtained from Dr. David Corney by personal communication.

Table 3: Precision comparison of IntEx and BioRAT from 229 abstracts.

Precision	IntEx		BioRAT	
Results	Cases	Percent (%)	Cases	Percent (%)
Correct	262	**65.66**	239	**55.07**
Incorrect	137	34.33	195	44.93
Totals	399	100.00	434	100.00

DIP contains protein interactions from both abstracts and full text. Since our extraction system was tested only on the abstracts, the system missed out on some interactions that were only present in the full text of the abstract.

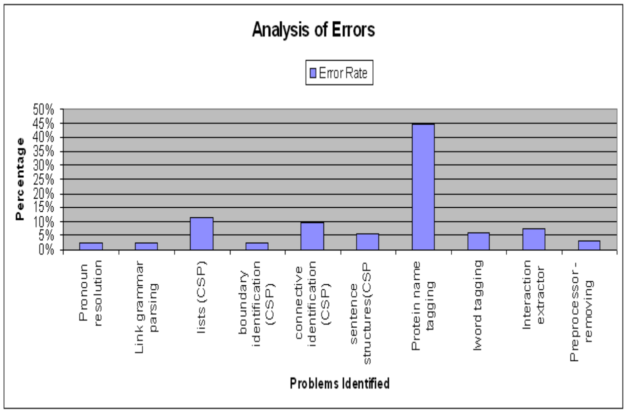


Figure 6: Analysis of types of errors encountered

Second evaluation for the IntEx system was done to test its recall performance using an article[6] that was also used by the GeneWays (Rzhetsky, Iossifov et al. 2004) system. Both systems performance was tested using the full text of the article (Friedman, Kra et al. 2001). GeneWays system achieves a recall of 65% where as IntEx extracted a total of 44 interactions corresponding to a recall measure of 66 %.

Conclusion

In this paper, we present a fully automated extraction system for identifying gene and protein interactions from biomedical text. The source code and documentation of the IntEx system, as well as all experimental documents and extracted interactions are available online at our Web site at http://cips.eas.asu.edu/textmining.htm. Our extraction system handles complex sentences and extracts multiple nested interactions specified in a sentence. Experimental evaluations of the IntEx system with the state of the art semi-automated systems -- the BioRAT and GeneWays datasets indicates that our system performs better without the labor intensive rule engineering requirement. We have shown that a syntactic role-based approach compounded with linguistically sound interpretation rules applied on the full sentence's parse can achieve better performance than existing systems which are based on manually engineered patterns which are both costly to develop and are not as scalable as the automated mechanisms presented in this paper.

Acknowledgements

We would like to thank to both Dr. David Corney and Dr. Andrew Rzhetsky for sharing their evaluation datasets and results.

References

Blaschke, C., M. A. Andrade, et al. (1999). "Automatic extraction of biological information from scientific text: Protein-protein interactions." Proceedings of International Symposium on Molecular Biology: 60-67.

Brill, E. (1992). "A simple rule-based part-of-speech tagger." Proceedings of ANLP 92, 3rd Conference on Applied Natural Language Processing: 152-155.

Bunescu, R., R. Ge, et al. (2003). Comparative Experiments on Learning Information Extractors for Proteins and their Interactions. Artificial Intelligence in Medicine.

Califf, M. E. and R. J. Mooney (1998). "Relational learning of pattern-match rules for information extraction." Working Notes of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing: 6-11.

Castaño, J., J. Zhang, et al. (2002). Anaphora Resolution in Biomedical Literature. International Symposium on Reference Resolution. Alicante, Spain.

Chiang, J.-H., H.-C. Yu, et al. (2004). "GIS: a biomedical text-mining system for gene information discovery." Bioinformatics 20(1): 120-121.

[6] Dataset was obtained from Dr. Andrew Rzhetsky by personal communication.

Corney, D. P. A., B. F. Buxton, et al. (2004). "BioRAT: extracting biological information from full-length papers." Bioinformatics 20(17): 3206-3213.

Craven, M. and J. Kumlien (1999). Constructing Biological Knowledge Bases by Extracting Information from Text Sources. Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology: 77--86.

Daelemans, W., S. Buchholz, et al. (1999). "Memory-based shallow parsing." Proceedings of CoNLL.

Ding, J., D. Berleant, et al. (2003). Extracting Biochemical Interactions from MEDLINE Using a Link Grammar Parser. Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03): 467.

Friedman, C., P. Kra, et al. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Proceedings of the International Confernce on Intelligent Systems for Molecular Biology: 574-82.

Fukuda, K., T. Tsunoda, et al. (1998). "Toward information extraction: Identifying protein names from biological papers." PSB 1998,: 705-716.

Grinberg, D., J. Lafferty, et al. (1995). "A Robust Parsing Algorithm For LINK Grammars." (CMU-CS-TR-95-125).

Hahn, U. and M. Romacker (2002). "Creating knowledge repositories from biomedical reports: The medsyndikate text mining system." Pacific Symposium on Biocomputing 2002: 338-349.

Hunter, R. T. a. C. R. a. J. (2000). "Extracting Molecular Binding Relationships from Biomedical Text." In Proceedings of the ANLP-NAACL 000,Association for Computational Linguistics: pages 188-195.

Leroy, G., H. Chen, et al. (2003). Genescene: biomedical text and data mining. Proceedings of the third ACM/IEEE-CS joint conference on Digital libraries: 116--118.

Mikheev, A. and S. Finch (1997). "A workbench for finding structure in texts." Proceedings of Applied Natural Language Processing (ANLP-97).

Novichkova, S., S. Egorov, et al. (2003). "MedScan, a natural language processing engine for MEDLINE abstracts." Bioinformatics 19(13): 1699-1706.

Ono, T., H. Hishigaki, et al. (2001). "Automatic Extraction of Information on protein-protein interactions from the biological literature." Bioinformatics 17(2): 155-161.

Phuong, T. M., D. Lee, et al. (2003). "Learning Rules to Extract Protein Interactions from Biomedical Text." PAKDD 2003: 148-158.

Porter, M. F. (1997). "An algorithm for suffix stripping." Progam, vol. 14, no. 3, July 1980: 313--316.

Quinlan, J. R. (1990). "Learning Logical Definitions from Relations." Mach. Learn. 5(3): 239--266.

Rinaldi, F., G. Schneider, et al. (2004). Mining relations in the GENIA corpus. Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics: 61 - 68.

Rzhetsky, A., I. Iossifov, et al. (2004). "GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data." J. of Biomedical Informatics 37(1): 43--53.

Seymore, K., A. McCallum, et al. (1999). Learning hidden markov model structure for information extraction. AAAI 99 Workshop on Machine Learning for Information Extraction.

Sleator, D. and D. Temperley (1991). Parsing English with a Link Grammar.Carnegie Mellon University Computer Science technical report CMU-CS-91-196, Carnegie Mellon University.

Sleator, D. and D. Temperley (1993). Parsing English with a Link Grammar. Third International Workshop on Parsing Technologies.

Szolovits, P. (2003). "Adding a Medical Lexicon to an English Parser." Proc. AMIA 2003 Annual Symposium.: 639-643.

Temkin, J. M. and M. R. Gilder (2003). "Extraction of protein interaction information from unstructured text using a context-free grammar." Bioinformatics 19(16): 2046-2053.

Thomas, J., D. Milward, et al. (2000). "Automatic extraction of protein interactions from scientific abstracts." Proceedings of the Pacific Symposium on Biocomputing 5: 502-513.

Thomas, R., C. Rajah, et al. (2000). "Extracting molecular binding relationships from biomedical text." Proceedings of the ANLP-NAACL 2000: 188-195.

Valencia, B. a. (2001). "Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study." Comp. Funct. Genom. 2: 196-206.

Xenarios, I., D. W. Rice, et al. (2000). "DIP: the Database of Interacting Proteins." Nucl. Acids Res. 28(1): 289-291.