Ontology-Based Natural Language Query Processing for the Biological Domain

Jisheng Liang, Thien Nguyen, Krzysztof Koperski, Giovanni Marchisio

Insightful Corporation
1700 Westlake Ave N., Suite 500, Seattle, WA, USA
{jliang,thien,krisk,giovanni}@insightful.com

Abstract

This paper describes a natural language query engine that enables users to search for entities, relationships, and events that are extracted from biological literature. The query interpretation is guided by a domain ontology, which provides a mapping between linguistic structures and domain conceptual relations. We focus on the usability of the natural language interface to users who are used to keyword-based information retrieval. Preliminary evaluation of our approach using the GENIA corpus and ontology shows promising results.

1 Introduction

New scientific research methods have greatly increased the volume of data available in the biological domain. A growing challenge for researchers and health care professionals is how to access this ever-increasing quantity of information [Hersh 2003]. The general public has even more trouble following current and potential applications. Part of the difficulty lies in the high degree of specialization of most resources. There is thus an urgent need for better access to current data and the various domains of expertise. Key considerations for improving information access include: 1) accessibility to different types of users; 2) high precision; 3) ease of use; 4) transparent retrieval across heterogeneous data sources; and 5) accommodation of rapid language change in the domain.

Natural language searching refers to approaches that enable users to express queries in explicit phrases, sentences, or questions. Current information retrieval engines typically return too many documents that a user has to go through. Natural language query allows users to express their information need in a more precise way and retrieve specific results instead of ranked documents. It also benefits users who are not familiar with domain terminology.

With the increasing availability of textual information related to biology, including MEDLINE abstracts and full-text journal articles, the field of biomedical text mining is rapidly growing. The application of Natural Language Processing (NLP) techniques in the biological domain has been focused on tagging entities, such as genes and proteins, and on detecting relations among those entities. The main goal of applying these techniques is database curation. There has been a lack of effort or success on improving search engine performance using NLP and text mining results. In this effort, we explore the feasibility of bridging the gap between text mining and search by

- Indexing entities and relationships extracted from text,
- Developing search operators on entities and relationships, and
- Transforming natural language queries to the entity-relationship search operators.

The first two steps are performed using our existing text analysis and search platform, called InFact [Liang 2005; Marchisio 2006]. This paper concerns mainly the step of NL query interpretation and translation. The processes described above are all guided by a domain ontology, which provides a conceptual mapping between linguistic structures and domain concepts/relations. A major drawback to existing NL query interfaces is that their linguistic and conceptual coverage is not clear to the user

Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 9–16,
New York City, June 2006. ©2006 Association for Computational Linguistics

[Androutsopoulos 1995]. Our approach addresses this problem by pointing out which concepts or syntactic relations are not mapped when we fail to find a consistent interpretation.

There has been skepticism about the usefulness of natural language queries for searching on the web or in the enterprise. Users usually prefer to enter the minimum number of words instead of lengthy grammatically-correct questions. We have developed a prototype system to deal with queries such as "With what genes does AP-1 interact?" The queries do not have to be standard grammatical questions, but rather have forms such as: "proteins regulated by IL-2" or "IL-2 inhibitors". We apply our system to a corpus of molecular biology literature, the GENIA corpus. Preliminary experimental results and evaluation are reported.

2 Overview of Our Approach

Molecular biology concerns interaction events between proteins, drugs, and other molecules. These events include transcription, translation, dissociation, etc. In addition to basic events which focus on interactions between molecules, users are also interested in relationships between basic events, e.g. the causality between two such events [Hirschman 2002]. In order to produce a useful NL query tool, we must be able to correctly interpret and answer typical queries in the domain, e.g.:

- What genes does transcription factor X regulate?
- With what genes does gene G physically interact?
- What proteins interact with drug D?
- What proteins affect the interaction of another protein with drug D?

Figure 1 shows the process diagram of our system. The query interpretation process consists of two major steps: 1) Syntactic analysis – parsing and decomposition of the input query; and 2) Semantic analysis – mapping of syntactic structures to an intermediate conceptual representation. The analysis uses an ontology to extract domain-specific entities/relations and to resolve linguistic ambiguity and variations. Then, the extracted semantic expression is transformed into an entity-relationship query language, which retrieves results from pre-indexed biological literature databases.

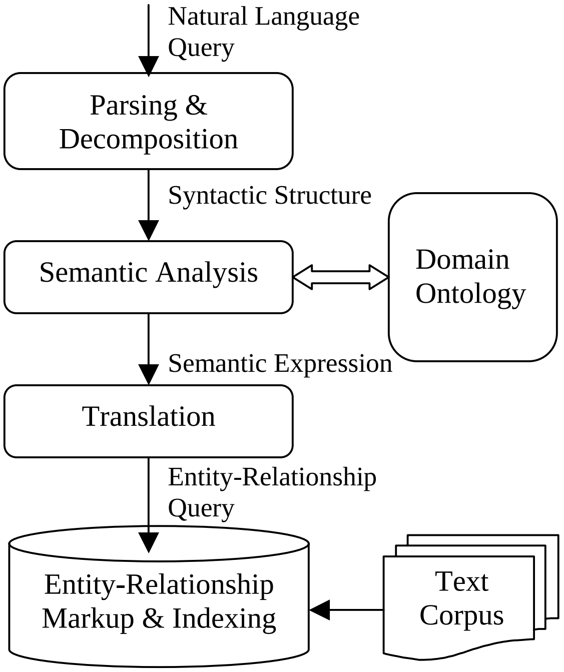


Figure 1 shows the query processing and retrieval process.

2.1 Incorporating Domain Ontology

Domain ontologies explicitly specify the meaning of and relation between the fundamental concepts in an application domain. A *concept* represents a set or class of entities within a domain. *Relations* describe the interactions between concepts or a concept's properties. Relations also fall into two broad categories: *taxonomies* that organize concepts into "is-a" and "is-a-member-of" hierarchy, and *associative* relationships [Stevens 2000]. The associative relationships represent, for example, the functions and processes a concept has or is involved in. A domain ontology also specifies how knowledge is related to linguistic structures such as grammars and lexicons. Therefore, it can be used by NLP to improve expressiveness and accuracy, and to resolve the ambiguity of NL queries.

There are two major steps for incorporating a domain ontology: 1) building/augmenting a lexicon for entity tagging, including lexical patterns that specify how to recognize the concept in text; and 2) specifying syntactic structure patterns for extracting semantic relationships among concepts. The existing ontologies (e.g. UMLS, Gene Ontology) are created mainly for the purpose of database annotation and consolidation. From those ontologies, we could extract concepts and taxonomic relations, e.g., is-a. However there is also a need for ontologies that specify relevant associative relations between concepts, e.g. “Protein acetylate Protein.” In our experiment we investigate the problem of augmenting an existing ontology (i.e. GENIA) with associative relations and other linguistic information required to guide the query interpretation process.

2.2 Query Parsing and Normalization

Our NL parser performs the steps of tokenization, part-of-speech tagging, morphological processing, lexical analysis, and identification of phrase and grammatical relations such as subjects and objects. The lexical analysis is based on a customizable lexicon and set of lexical patterns, providing the abilities to add words or phrases as dictionary terms, to assign categories (e.g. entity types), and to associate synonyms and related terms with dictionary items. The output of our parser is a dependency tree, represented by a set of dependency relationships of the form (head, relation, modifier).

In the next step, we perform syntactic decomposition to collapse the dependency tree into subject-verb-object (SVO) expressions. The SVO triples can express most types of syntactic relations between various entities within a sentence. Another advantage of this triple expression is that it becomes easier to write explicit transformational rules that encode specific linguistic variations.

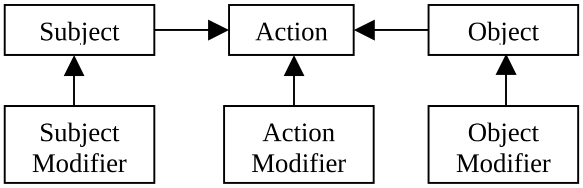


Figure 2 shows the subject-action-object triplet.

Verb modifiers in the syntactic structure may include prepositional attachment and adverbials. The modifiers add context to the event of the verb, including time, location, negation, etc. Subject/object modifiers include appositive, nominative, genitive, prepositional, descriptive (adjective-noun modification), etc. All these modifiers can be either considered as descriptors (attributes) or reformulated as triple expressions by assigning a type to the pair.

Linguistic normalization is a process by which linguistic variants that contain the same semantic content are mapped onto the same representational structure. It operates at the morphological, lexical and syntactic levels. Syntactic normalization involves transformational rules that recognize the equivalence of different structures, e.g.:

- Verb Phrase Normalization – elimination of tense, modality and voice.
- Verbalization of noun phrases – e.g. Inhibition of X by Y → Y inhibit X.

For example, queries such as:

Proteins activated by IL-2
What proteins are activated by IL-2?
What proteins does IL-2 activate?
Find proteins that are activated by IL-2

are all normalized into the relationship:

IL-2 > activate > Protein

As part of the syntactic analysis, we also need to catch certain question-specific patterns or phrases based on their part-of-speech tags and grammatical roles, e.g. determiners like “which” or “what”, and verbs like “find” or “list”.

2.3 Semantic Analysis

The semantic analysis typically involves two steps: 1) Identifying the semantic type of the entity sought by the question; and 2) Determining additional constraints by identifying relations that ought to hold between a candidate answer entity and other entities or events mentioned in the query [Hirschman 2001]. The semantic analysis attempts to map normalized syntactic structures to semantic entities/relations defined in the ontology. When the system is not able to understand the question, the cause of failure will be explained to the user, e.g. unknown word or syntax, no relevant concepts in the ontology, etc. The output of semantic analysis is a set of relationship triplets, which can be grouped into four categories:

Events, including interactions between entities and inter-event relations (nested events), e.g.

Inhibition(“il-2”, “erbb2”)
Inhibition(protein, Activation(DEX, IkappaB))

Event Attributes, including attributes of an interaction event, e.g.

Location(Inhibition(il-2, erbb2), “blood cell”)

Entity Attributes, including attributes of a given entity, e.g.

Has-Location(“erbb2”, “human”)

Entity Types, including taxonomic paths of a given entity, e.g.

Is-A(“erbb2”, “Protein”)

A natural language query will be decomposed into a list of inter-linked triplets. A user’s specific information request is noted as “UNKNOWN.”

Starting with an ontology, we determine the mapping from syntactic structures to semantic relations. Given our example “IL-2 > activate > Protein”, we recognize “IL-2” as an entity, map the verb “activate” to a semantic relation “Activation,” and detect the term “protein” as a designator of the semantic type “Protein.” Therefore, we could easily transform the query to the following triplets:

- Activation(IL-2, UNKNOWN)
- Is-A(UNKNOWN, Protein)

Given a syntactic triplet of subject/verb/object or head/relation/modifier, the ontology-driven semantic analysis performs the following steps:

1. Assign possible semantic types to the pair of terms,
2. Determine all possible semantic links between each pair of assigned semantic types defined in the ontology,
3. Given the syntactic relation (i.e. verb or modifier-relation) between the two concepts, infer and validate plausible inter-concept semantic relationships from the set determined in Step 2,
4. Resolve linguistic ambiguity by rejecting inconsistent relations or semantic types.

It is simpler and more robust to identify the query pattern using the extracted syntactic structure, in which linguistic variations have been normalized into a canonical form, rather than the original question or its full parse tree.

2.4 Entity-Relationship Indexing and Search

In this section, we describe the annotation, indexing and search of text data. In the off-line indexing mode, we annotate the text with ontological concepts and relationships. We perform full linguistic analysis on each document, which involves splitting of text into sentences, sentence parsing, and the same syntactic and semantic analysis as described in previous sections on query processing. This step recognizes names of proteins, drugs, and other biological entities mentioned in the texts. Then we apply a document-level discourse analysis procedure to resolve entity-level coreference, such as acronyms/aliases and pronoun anaphora. Sentence-level syntactic structures (subject-verb-object triples) and semantic markups are stored in a database and indexed for efficient retrieval.

In the on-line search mode, we provide a set of entity-relationship (ER) search operators that allow users to search on the indexed annotations. Unlike keyword search engines, we employ a highly expressive query language that combines the power of grammatical roles with the flexibility of Boolean operators, and allows users to search for actions, entities, relationships, and events. We represent the basic relationship between two entities with an expression of the kind:

Subject Entity > Action > Object Entity

We can optionally constrain this expression by specifying modifiers or using Boolean logic. The arrows in the query refer to the directionality of the action. For example,

Entity 1 <> Action <> Entity 2

will retrieve all relationships involving *Entity 1* and *Entity 2*, regardless of their roles as subject or object of the action. An asterisk (*) can be used to denote unknown or unspecified sources or targets, e.g. “Il-2 > inhibit > *”.

In the ER query language we can represent and organize entity types using taxonomy paths, e.g.:

[substance/compound/amino_acid/protein]
[source/natural/cell_type]

The taxonomic paths can encode the “is-a” relation (as in the above examples), or any other relations defined in a particular ontology (e.g. the “part-of” relation). When querying, we can use a taxonomy path to specify an entity type, e.g. *[Protein/Molecule]*, *[Source]*, and the entity type will automatically include all subpaths in the taxonomic hierarchy. The complete list of ER query features that we currently support is given in Table 1.

ER Query Features	Descriptions and Examples
Relationships between two entities or entity types	The query “*il-2 <> * <> Ap1*” will retrieve all relationships between the two entities.
Events involving one or more entities or types	The query “*il-2 > regulate > [Protein]*” will return all instances of il-2 regulating a protein.
Events restricted to a certain action type - categories of actions that can be used to filter or expand search	The query “*[Protein] > [Inhibition] > [Protein]*” will retrieve all events involving two proteins that are in the nature of inhibition.
Boolean Operators - AND, OR, NOT	Example: Il-2 OR “interleukin 2” > inhibit or suppress >* Phrases such as “interleukin 2” can be included in quotes.
Prepositional Constraints - Filter results by information found in a prepositional modifier.	Query *Il-2 > activate > [protein]^[cell_type]* will only return results mentioning a cell type location where the activation occurs.
Local context constraints - Certain keyword(s) must appear near the relationship (within one sentence).	Example: LPS > induce > NF-kappaB CONTEXT CONTAINS “human T cell”
Document keyword constraints - Documents must contain certain keyword(s)	Example: Alpha-lipoic acid > inhibit > activation DOC CONTAINS “AIDS” OR “HIV”
Document metadata constraints	Restrict results to documents that contain the specified metadata values.
Nested Search	Allow users to search the results of a given search.
Negation Filtering	Allow users to filter out negated results that are detected during indexing.

Table 1 lists various types of ER queries

2.5 Translation to ER Query

We extract answers through entity-relational matching between the NL query and syntactic/semantic annotations extracted from sentences. Given the query’s semantic expression as described in Section 2.3, we translate it to one or more entity-relationship search operators. The different types of semantic triplets (i.e. Event, Attribute, and Type) are treated differently when being converted to ER queries.

- The Event relations can be converted directly to the subject-action-object queries.
- The inter-event relations are represented as local context constraints.
- The Event Attributes are translated to prepositional constraints.
- The Entity Attribute relations could be extracted either from same sentence or from somewhere else within document context, using the nested search feature.
- The Entity Type relations are specified in the ontology taxonomy.

For our example, “proteins activated by il-2”, we translate it into an ER query: “il-2 > [activation] > [protein]”. Figure 3 shows the list of retrieved subject-verb-object triples that match the query, where each triple is linked to a sentence in the corpus.

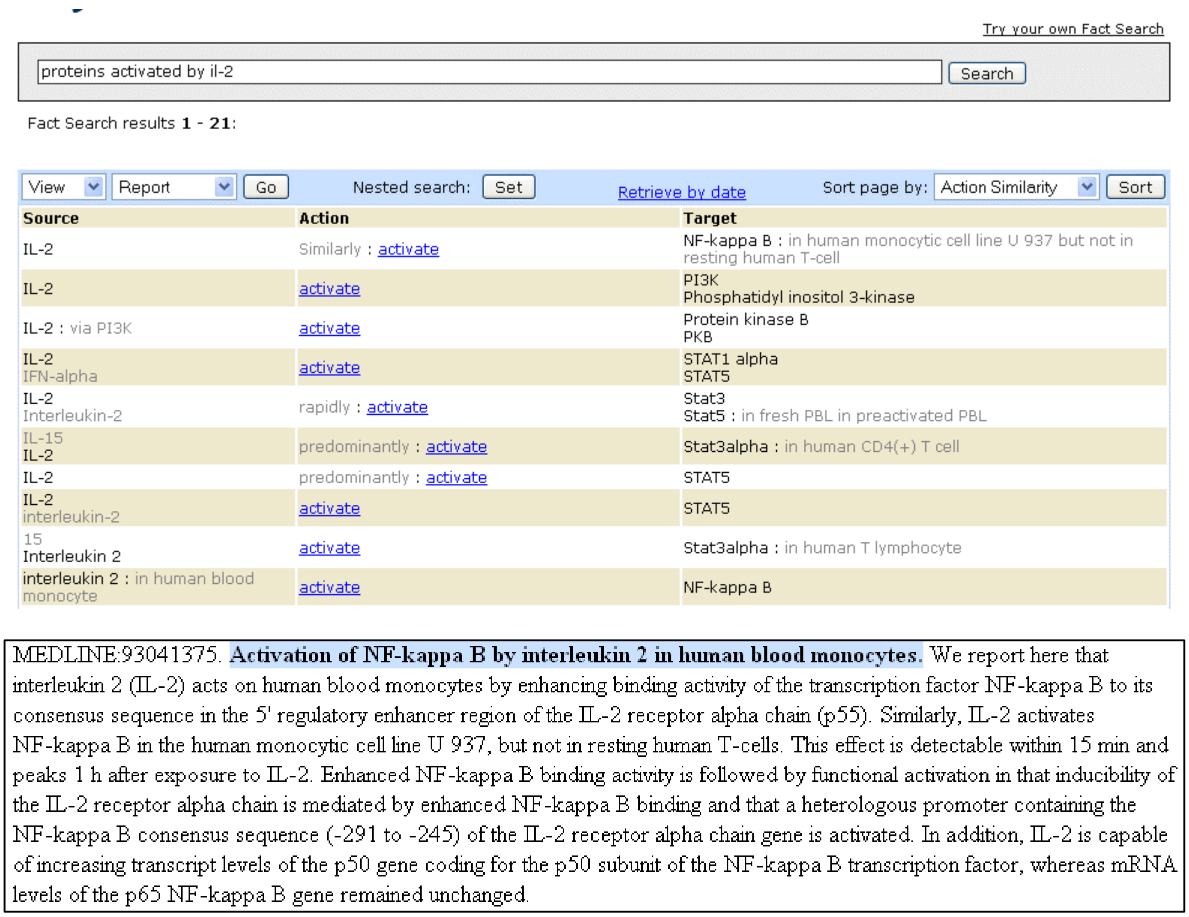
3 Experiment Results

We tested our approach on the GENIA corpus and ontology. The evaluation presented in this section focuses on the ability of the system to translate NL queries into their normalized representation, and the corresponding ER queries.

3.1 Test Data

The GENIA corpus contains 2000 annotated MEDLINE abstracts [Ohta 2002]. The main reason we chose this corpus is that we could extract the pre-annotated biological entities to populate a domain lexicon, which is used by the NL parser. Therefore, we were able to ensure that the system had complete terminology coverage of the corpus. During indexing, we used the raw text data as input by stripping out the annotation tags.

The GENIA ontology has a complete taxonomy of entities in molecular biology. It is divided into substance and source sub-hierarchies. The substances include sub-paths such as nucleic_acid/DNA and amino_acid/protein. Sources are biological locations where substances are found and their reactions take place. They are also hierarchically subclassified into organisms, body parts, tissues, cells



proteins activated by il-2

Fact Search results **1** - **21**:

Source	Action	Target
IL-2	Similarly : activate	NF-kappa B : in human monocytic cell line U 937 but not in resting human T-cell
IL-2	activate	PI3K Phosphatidyl inositol 3-kinase
IL-2 : via PI3K	activate	Protein kinase B PKB
IL-2 IFN-alpha	activate	STAT1 alpha STAT5
IL-2 Interleukin-2	rapidly : activate	Stat3 Stat5 : in fresh PBL in preactivated PBL
IL-15 IL-2	predominantly : activate	Stat3alpha : in human CD4(+) T cell
IL-2	predominantly : activate	STAT5
IL-2 interleukin-2	activate	STAT5
15 Interleukin 2	activate	Stat3alpha : in human T lymphocyte
interleukin 2 : in human blood monocyte	activate	NF-kappa B

MEDLINE:93041375. **Activation of NF-kappa B by interleukin 2 in human blood monocytes.** We report here that interleukin 2 (IL-2) acts on human blood monocytes by enhancing binding activity of the transcription factor NF-kappa B to its consensus sequence in the 5' regulatory enhancer region of the IL-2 receptor alpha chain (p55). Similarly, IL-2 activates NF-kappa B in the human monocytic cell line U 937, but not in resting human T-cells. This effect is detectable within 15 min and peaks 1 h after exposure to IL-2. Enhanced NF-kappa B binding activity is followed by functional activation in that inducibility of the IL-2 receptor alpha chain is mediated by enhanced NF-kappa B binding and that a heterologous promoter containing the NF-kappa B consensus sequence (-291 to -245) of the IL-2 receptor alpha chain gene is activated. In addition, IL-2 is capable of increasing transcript levels of the p50 gene coding for the p50 subunit of the NF-kappa B transcription factor, whereas mRNA levels of the p65 NF-kappa B gene remained unchanged.

Figure 3 shows our natural language query interface. The retrieved subject-verb-object relationships are displayed in a tabular format. The lower screenshot shows the document display page when user clicks on the last result link <interleukin 2, activate, NF-kappa B>. The sentence that contains the result relationship is highlighted.

or cell types, etc. Our adoption of the GENIA ontology as a conceptual model for guiding query interpretation is described as follows.

Entities - For gene and protein names, we added synonyms and variations extracted from the Entrez Gene database (previously LocusLink).

Interactions – The GENIA ontology does not contain associative relations. By consulting a domain expert, we identified a set of relations that are of particular interest in this domain. Some examples of relevant relations are: activate, bind, interact, regulate. For each type of interaction, we created a list of corresponding action verbs.

Entity Attributes - We identified two types of entity attributes:

1. Location, e.g. body_part, cell_type, etc. identified by path [genia/source]
2. Subtype of proteins/genes, e.g. enzymes, transcription factors, etc., identified by types like protein_family_or_group, DNA_family_or_group

Event Attributes - Locations were the only event attribute we supported in this experiment.

Designators - We added a mapping between each semantic type and its natural language names. For example, when a term such as "gene" or "nucleic acid" appears in a query, we map it to the taxonomic path: [Substance/compound/nucleic_acid]

3.2 Evaluation

To demonstrate our ability to interpret and answer NL queries correctly, we selected a set of 50 natural language questions in the molecular biology domain. The queries were collected by consulting a domain expert, with restrictions such as:

1. Focusing on queries concerning entities and interaction events between entities.
2. Limiting to taxonomic paths defined within the GENIA ontology, which does not contain important entities such as drugs and diseases.

For each target question, we first manually created the ground-truth entity-relationship model. Then, we performed automatic question interpretation and answer retrieval using the developed software prototype. The extracted semantic expressions were verified and validated by comparison against the ground-truth. Our system was able to correctly interpret all the 50 queries and retrieve answers from the GENIA corpus. In the rest of this section, we describe a number of representative queries.

Query on events:
With what genes does ap-1 physically interact?
Relations:
Interaction(“ap-1”, UNKOWN)
IS-A(UNKNOWN, “Gene”)
ER Query:
ap-1 <>[Interaction] <> [nucleic_acid]

Queries on association:
erbb2 and il-2
what is the relation between erbb2 and il-2?
Relations:
Association(“erbb2”, “il-2”)
ER Query:
Erbb2 <>*<>il-2

Query of noun phrases:
Inhibitor of erbb2
Relation:
Inhibition(UNKNOWN, “erbb2”)
ER Query:
[substance] > [Inhibition] > erbb2

Query on event location:
In what cell types is il-2 activated?
Relations:
Activation (*, “Il-2”)
Location (Activation(), [cell_type])
ER Query:
* > [Activation] > il-2 ^ [cell_type]

Entity Attribute Constraints

An entity’s properties are often mentioned in a separate place within the document. We translate these types of queries into DOC_LEVEL_AND of multiple ER queries. This AND operator is currently implemented using the feature of nested search. For example, given query**:**

What enzymes does HIV-1 Tat suppress?

we recognize the word "enzyme" is associated with the path: [protein/protein_family_or_group], and we consider it as an attribute constraint.

Relations:
Inhibition (“hiv-1 tat”, UNKNOWN)
IS-A(UNKNOWN, “Protein”)
HAS-ATTRIBUTE (UNKNOWN, “enzyme”)
ER query:
(hiv-1 tat > [Inhibition]> [protein])
DOC_LEVEL_AND
([protein] > be > enzyme)

One of the answer sentences is displayed below:

> “Thus, our experiments demonstrate that the C-terminal region of HIV-1 Tat is required to suppress Mn-SOD expression”

while Mn-SOD is indicated as an enzyme in a different sentence:

> “… Mn-dependent superoxide dismutase (**Mn-SOD**), a mitochondrial **enzyme** … ”

Inter-Event Relations

The inter-event relations or nested event queries (CLAUSE_LEVEL_AND) are currently implemented using the ER query’s local context constraints, i.e. one event must appear within the local context of the other.

Query on inter-event relations:
What protein inhibits the induction of Ikappa-Balpha by DEX?
Relations:
Inhibition ([protein], Activation())
Activation (“DEX”, “IkappaBalpha”)
ER Query:
([protein] > [Inhibition] > *)
CLAUSE_LEVEL_AND
(DEX > [Activation] > IkappaBalpha)

One of the answer sentences is:

> "In both cell types, the cytokine that inhibits the induction of IkappaBapha by DEX, also rescues these cells from DEX-induced apoptosis."

4 Discussions

We demonstrated the feasibility of our approach using the relatively small GENIA corpus and ontology. A key concern with knowledge or semantic based methods is the scalability of the methods to larger set of data and queries. As future work, we plan to systematically measure the effectiveness of the approach based on large-scale experiments in an information retrieval setting, as we increase the knowledge and linguistic coverage of our system.

We are able to address the large data size issue by using InFact as an ingestion and deployment platform. With a distributed architecture, InFact is capable of ingesting large data sets (i.e. millions of MEDLINE abstracts) and hosting web-based search services with a large number of users. We will investigate the scalability to larger knowledge coverage by adopting a more comprehensive ontology (i.e. UMLS [Bodenreider 2004]). In addition to genes and proteins, we will include other entity types such as drugs, chemical compounds, diseases and phenotypes, molecular functions, and biological processes, etc. A main challenge will be increasing the linguistic coverage of our system in an automatic or semi-automatic way.

Another challenge is to encourage keyword search users to use the new NL query format and the semi-structured ER query form. We are investigating a number of usability enhancements, where the majority of them have been implemented and are being tested.

For each entity detected within a query, we provide a hyperlink that takes the user to an ontology lookup page. For example, if the user enters "protein il-2", we let the user know that we recognize "protein" as a taxonomic path and "il-2" as an entity according to the ontology. If a relationship triplet has any unspecified component, we provide recommendations (or tips) that are hyperlinks to executable ER queries. This allows users who are not familiar with the underlying ontology to navigate through most plausible results. When the user enters a single entity of a particular type, we display a list of relations the entity type is likely to be involved in, and a list of other entity types that are usually associated to the given type. Similarly, we define a list of relations between each pair of entity types according to the ontology. The relations are ranked according to popularity. When the user enters a query that involves two entities, we present the list of relevant relations to the user.

Acknowledgements: This research was supported in part by grant number 1 R43 LM008464-01 from the NIH. The authors thank Dr. David Haynor for his advice on this work; the anonymous reviewers for their helpful comments; and Yvonne Lam for helping with the manuscript.

References

Androutsopoulos I, Ritchie GD and Thanisch P. "Natural Language Interfaces to Databases – An Introduction", *Journal of Natural Language Engineering,* Vol 1, pp. 29-81, 1995.

Bodenreider O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research,* 2004**.**

Hersh W and Bhupatiraju RT. "TREC Genomics Track Overview", In *Proc. TREC,* 2003, pp. 14-23.

Hirschman L and Gaizauskas R. Natural Language Question Answering: The View from Here. *Natural Language Engineering*, 2001.

Hirschman L, Park JC, Tsujii J, Wong L and Wu CH. Accomplishments and Challenges in Literature Data Mining for Biology. *Bioinformatics Review*, Vol. 18, No. 12, 2002, pp. 1553-1561.

Liang J, Koperski K, Nguyen T, and Marchisio G. Extracting Statistical Data Frames from Text. *ACM SIGKDD Explorations*, Volume 7, Issue 1, pp. 67 – 75, June 2005.

Marchisio G, Dhillon D, Liang J, Tusk C, Koperski K, Nguyen T, White D, and Pochman L. A Case Study in Natural Language Based Web Search. To appear in *Text Mining and Natural Language Processing*. A Kao and SR Poteet (Editors). Springer 2006.

Ohta T, Tateisi Y, Mima H, and Tsujii J. GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. In *Proc. HLT* 2002.

Stevens R, Goble CA, and Bechhofer S. Ontology-based Knowledge Representation for Bioinformatics. *Briefings in Bioinformatics*, November 2000.