RC24093 (W0610-164) October 30, 2006
Computer Science

IBM Research Report

Discourse Semantics for Biomedical Information Discovery

Arendse Bernth
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Discourse Semantics for Biomedical Information Discovery

Arendse Bernth
IBM T.J. Watson Research Center
19 Skyline Dr., Hawthorne, NY 10532, USA
arendse@us.ibm.com

Abstract

We describe a computational system for representing and extracting the meaning of a natural language discourse, with an application of the system to information discovery in the biomedical domain.

1 Introduction

Euphoria is a system that constructs a discourse-level semantic analysis based on deep sentence-level parsing and discourse-semantic processing including coreference, resolution of implicit arguments, and most-plausible semantics. The result is production of a normalized semantic representation, conveniently stored in a database indexed by *extended entities*. Whereas the system is quite general, it has recently been applied to information discovery in the medical domain.

Section 2 gives an overview of how Euphoria constructs the meaning representation. In section 3 a brief description of the meaning representation formalism is given, and examples in section 4 illustrate the use of Euphoria for information discovery in the biomedical domain.

2 Producing the semantic representation

The first step in producing the semantic representation is deep syntactic parsing on a sentence level. For this, we use the English Slot Grammar [McC80, McC90, McC06]. This is a broad-coverage, general parser that gives us information about both near and long distance relations, and in some cases also fills in implicit arguments.

The second step is application of a mix of sentence and discourse level analysis. First, unique referent IDs, either new or existing, are assigned to most words in the sentence as a result of coreference resolution.

Then the parse tree is explored to determine the relations between constituents of the sentence in order to produce the semantic representation. However, there is by no means a one-one correspondence between syntactic units and the semantic representation. This is due to two things. First, the semantic representation is a *normalized* representation, where *e.g.* active and passive sentences are given the same semantic interpretation.[1] Secondly, some referents may need to be made explicit. For example, in the case of ellipsis, Euphoria attempts to identify the missing constiuents and produces a semantic representation that is fuller than what is actually present in the surface structure.

In case of real syntactic ambiguity requiring domain knowledge to resolve it, such as attachment of present participles following the object that could be attached either to the subject or the object, selectional preferences are applied to determine the attachment. (See [BM03].)

During the exploration of the parse tree, implicit arguments are identified and filled in. Some of the implicit arguments are determined during parsing; others are handled during the production of the semantic representation. (See [Ber06].)

3 Semantic representation

The resulting semantic representation, described more fully in [Ber06], is a flat semantic structure expressed in terms of *entity-oriented logical forms* (*EOLF*s), which make use of *extended entities* (*EE*s). EEs include basically anything that can be referred to, such as events and relations in addition to entities in the conventional sense.

The EOLF formalism expresses the referability of all entities by using a generalization of the indexing of verbs by so-called event variables first proposed by [Dav67]. Our "events" are indeed very general, along the lines described in [Hob85] and [MB05].

Such generalized "events" furthermore have the advantage of allowing a flat semantic structure, a property that makes automatic reasoning easier. Flat structures are also used in [Hob85], but there are differences.

An *entity-oriented logical form* (EOLF) consists of an extended entity E (called the *index* of the EOLF), together with a set S of *predications* that are "about" E, in the sense that E appears in each member of S. In the examples, we will display the EOLFs as follows:

[1] We actually think that two different surface structures never express *exactly* the same proposition; however, this view does not seem viable for a practical application.

`Index` < `(Predication`$_1$ `...` `Predication`$_n$`)`.
These EOLFs are stored in a database indexed by the `Index`.

4 Application to the biomedical domain

In this section we show some examples from the biomedical domain with some interesting relations extracted by Euphoria. Extracting relations is a crucial first step in information discovery. For ease of reading, the extracted relations in (1) are glossed in English by our generation module. We give only the relevant parts of the database, but the generation module has access to the complete database, hence is able to include some modifiers, etc., not present in the relations shown here.

Example (1) (from [YTMT01]) illustrates the value of deep parsing and analysis.

(1) a. An active phorbol ester must therefore presumably by activation of protein kinase cause dissociation of a cytoplasmic complex of NF-kappa B and I kappa B by modifying I kappa B.

b. `((activate#7V Phorbol_Ester#5 protein_kinase#9))`
An phorbol ester activates the protein kinase.

c. `((modify#25V Phorbol_Ester#5 i_kappa_b#21))`
The phorbol ester modifies an I kappa B.

d. `((dissociate#11V Phorbol_Ester#5 complex#13))`
The phorbol ester dissociates a cytoplasmic complex.

e. `((cause#23V Phorbol_Ester#5 dissociate#11V))`
The phorbol ester causes the cytoplasmic complex to dissociate.

The relation shown in (1b) illustrates the resolution of the implicit argument of `activation`, a relation that has been normalized to the verbal form `activate`. In (1c) we show the impact of deep parsing to capture the long-distance relation that `phorbol ester` exhibits as the implicit subject of `modifying`. The case in (1d) is similar to (1b); and in (1e) the implicit, and distant, subject of `cause` is resolved.

Example (2) shows the combined effect of coreference resolution and determination of implicit arguments. Deep parsing identifies the implicit subject of `protect` as `it`, and the coreference module has resolved the referent of `it` to `Nucleophosmin`.

(2) a. **Nucleophosmin** (NPM) is a multifunctional protein frequently overexpressed in actively proliferating cells including tumor and stem cells. Here we show that **it** acts as a cellular p53 negative regulator to **protect** normal and malignant hematopoietic cells from stress-induced apoptosis.

b. `protect#50V < ((protect#50V Nucleophosmin#3 cell#16G from#59P))`

References

[Ber06] Arendse Bernth. Implicit predicate arguments and discourse. Technical Report RC 24046, 2006. Submitted to the special issue on Computational Approaches to Discourse and Document Processing in the journal Traitement Automatique des Langues.

[BM03] Arendse Bernth and Michael C. McCord. A hybrid approach to deriving selectional preferences. In *Proceedings of MT Summit IX*, pages 9–15, New Orleans, 2003.

[Dav67] Donald Davidson. The logical form of action sentences. In N. Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press, Pittsburgh, Pa., 1967.

[Hob85] Jerry R. Hobbs. Ontological promiscuity. In *Proc. of the 23rd ACL*, pages 61–69, Chicago, IL, 1985.

[MB05] Michael C. McCord and Arendse Bernth. A metalogical theory of natural language semantics. *Linguistics and Philosophy*, 28:73–116, 2005.

[McC80] Michael C. McCord. Slot Grammars. *Computational Linguistics*, 6:31–43, 1980.

[McC90] Michael C. McCord. Slot Grammar: A system for simpler construction of practical natural language grammars. In R. Studer, editor, *Natural Language and Logic: International Scientific Symposium, Lecture Notes in Computer Science*, pages 118–145. Springer Verlag, Berlin, 1990.

[McC06] Michael C. McCord. A formal system for Slot Grammar. Technical report, IBM T. J. Watson Research Center, 2006. RC 23976.

[YTMT01] Akane Yakushiji, Yuka Tateisi, Yusuke Miyao, and Jun-Ichi Tsujii. Event extraction from biomedical papers using a full parser. In *Pacific Symposium on Biocomputing 6*, pages 408–419, Hawaii, 2001.