Available online at www.sciencedirect.com
SCIENCE DIRECT®
Data & Knowledge Engineering 55 (2005) 59–83
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

Ontology-driven discourse analysis for information extraction

Philipp Cimiano [a,*], Uwe Reyle [b], Jasmin Šarić [c]

[a] Institute AIFB, University of Karlsruhe, Englerstr. 11, 76131 Karlsruhe, Germany
[b] Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Azenbergstr. 12, 70174 Stuttgart, Germany
[c] EML Research gGmbH, Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany

Received 4 November 2004; accepted 4 November 2004
Available online 21 December 2004

Abstract

This paper presents a novel approach to discourse analysis within information extraction systems. It makes use of DRT as formal representation of the linguistic context as well as of a domain-specific ontology as a basis to compute conceptual relations between extracted events thus establishing discourse coherence. The approach has been implemented within GenIE, an information extraction system with the aim of extracting information about biochemical pathways, about sequences, structures and functions of genomes and proteins. The approach is evaluated against a semantically hand-annotated set of Swiss-Prot protein function descriptions and shows very promising results.

Keywords: Information extraction; Discourse analysis; Event ontology; Biomedical NLP

1. Introduction

The most important source of biochemical data is the fast growing number of articles available in electronic form. Medline [1] for example contains over 15 million abstracts and approximately

* Corresponding author.
E-mail addresses: cimiano@aifb.uni-karlsruhe.de (P. Cimiano), uwe@ims.uni-stuttgart.de (U. Reyle), Jasmin.Saric@eml-r.villa-bosch.de (J. Šarić).
[1] http://www.ncbi.nlm.nih.gov/PubMed/.

0169-023X/$ - see front matter
doi:10.1016/j.datak.2004.11.009

40,000 are added each month. Other important resources are the Journal of Biological Chemistry [2] with more than 50,000 pages published per year as well as the Swiss-Prot database, [3] which contains natural language descriptions of the function of each protein. This huge amount of unstructured information has in fact become to be known as the 'biobibliome'. Indeed, it seems crucial to exploit natural language processing techniques to extract information from these free text sources, structure it and feed databases with it. The storage and organization of this biochemical knowledge in a database can in turn facilitate the reasoning about the data and lead to the understanding of specific biochemical processes as well as to the discovery of new aspects of them.

Information Extraction (IE) is the task of identifying, collecting and normalizing relevant information from natural language texts and producing a set of target knowledge structures as output. These target knowledge structures are defined by a given ontology which represents a model of the domain in question and thus also specifies which information is relevant. In fact, a lot of research in IE is concentrating on biomedical or biochemical articles as domains of application. Some researches have focused in particular on the extraction of events, i.e. the dynamic aspects of the domain in question [1–5].

However, most state-of-the-art information extraction systems in the biochemical domain are limited to the extraction of isolated events without situating them properly within the context of other extracted events. The following two examples taken from the Swiss-Prot database clearly show the necessity to establish contextual dependencies between events:

(1) U1 SNRNP A BINDS STEM LOOP II OF U1 SNRNA. [...]
THIS INTERACTION IS REQUIRED FOR THE SUBSEQUENT BINDING OF U2 SN-RNP AND THE U4/U6/U5 TRI-SN-RNP.

(2) (TMF) THIS PROTEIN BINDS THE HIV-1 TATA ELEMENT AND INHIBITS TRANSCRIPTIONAL ACTIVATION BY THE TATA-BINDING PROTEIN (TBP).

In the first example, it is important to resolve the definite description 'THIS INTERACTION' as referring to the binding event mentioned in the first sentence. Only then will we get the correct interpretation that the binding event of the first sentence is the one 'REQUIRED FOR THE SUBSEQUENT BINDING' mentioned in the second one.

In the second example, it is clearly not enough to extract the *bind* and *inhibit* events in isolation. Only if we identify that the relation between the extracted events is a resultative one, will we yield the correct interpretation of the sentence, i.e. that it is the binding of TMF to the HIV-1 TATA element which inhibits the transcriptional activation by TBP.

It has become clear that it is not enough to extract isolated events but that they have to be embedded within the context they are extracted from. Thus, the necessity of a linguistic approach which identifies conceptual relations between extracted events seems obvious. On the other hand, information extraction systems are typically restricted to a specific domain of application so that it seems feasible to create a conceptual model of the domain which can be exploited within such an approach.

[2] http://www.jbc.org.

[3] At the time of writing it contained over 159,000 protein sequence entries (http://www.expasy.org/sprot/).

This paper presents a knowledge-based approach to discourse analysis which on the basis of a given ontology and a semantic representation of events extracted from the text computes relations between them that are predefined in the ontology representing a model of the domain. The structure of the paper is as follows: Section 2 presents the overall architecture of the system and briefly describes each of its components. Section 3 presents a corpus study which motivates the necessity of a discourse analysis component from a quantitative point of view and Section 4 describes the underlying ontological model as well as the principles according to which an appropriate ontology of biochemical events has been developed. Section 5 presents the ontology-driven approach to discourse analysis and Section 6 presents the evaluation of the discourse analysis component against a set of short texts describing the function of proteins taken from the Swiss-Prot database. Finally, Section 7 discusses related work and Section 8 concludes the paper.

2. The GenIE system

GenIE (Genome Information Extraction) is a system with the aim of extracting information about biochemical pathways, about sequences, structures and functions of genomes and proteins from biomedical literature and database comment lines. For this purpose, it makes use of both shallow processing techniques and classical linguistic approaches based on deep syntactic and semantic analysis guided by ontological information.

The processing architecture of GenIE differs crucially from the standard linear architecture of Information Extraction systems, as it is shown in Fig. 1. These systems typically consist of a cascade of transducers which transform representations of one intermediate level into representations of the next level. The first step requires the determination of sentence- and token-boundaries as well as the recognition of multi-word terms. Then structure is added by associating part of speech (POS) tags and lexical attributes to words which then are combined to small-scale syntactic structures (chunks) such as noun and verb groups and other phrases that can be recognized with high reliability. Recognized words and phrases are translated into semantic forms which are then composed into more complex representations according to their predicate/argument structure which in their turn are combined by merging techniques (compare [6]) to the final representations.

There is more than one way in which this processing architecture is too simple. First, the linear architecture of these systems often leads to incorrect analyses because any mistake made at an earlier level will propagate itself through the whole cascade. [4] Second, there is no leeway to operate at and move between levels of varying depth of analysis. The architecture in Fig. 1 is based on the assumption that shallow representations (e.g. a set of syntactic chunks) may be all that is required as input for the computation of a final representation (a full semantic representation, or a template), because the restrictions of the final level permit only one coherent way in which the shallow representation bits may be combined. But most of the time this assumption is not met by the input to be processed. On the other hand, deep analysis might be required at any level of the cascade at least for some parts of their input. The recognition of names for biochemical molecules, enzymes, protein complexes or genes for example might require a deep analysis (involving morphological,

[4] Hobbs et al. [6] report that many of the cases where information provided by some component would be most useful are precisely those in which the component is most likely to make errors.

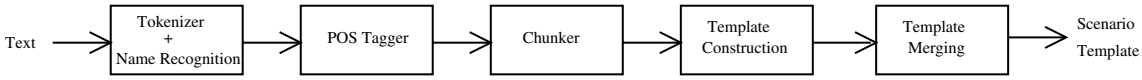

Fig. 1. Linear IE architecture.

syntactical and semantical interpretations at—again—varying depth) in case these names are not listed in the lexicon and cannot be reliably dealt with by simple heuristics. Another example is the resolution of coreferences. As will be shown in the later sections of this paper a proper treatment of these phenomena requires the interaction between information projected from the semantic lexicon, the ontology and the discourse analysis component.

GenIE accounts for these requirements by a system architecture that supports the information flow between different levels and depths of representation as shown in Fig. 2. Along the vertical axis on the left we have again plotted what are commonly regarded as the main components of Information Extraction systems. Depth of analysis is gained when the system includes components depicted along the axis on the right hand side of the picture. The central axis of the schema shows the lexical and other knowledge sources needed for the particular analysis tasks (indicated by dashed arrows).

The semantic representation formalism used in GenIE is called Discourse Representation Theory (DRT, [7]). The language of DRT (the so-called Discourse Representation Structures, DRSs) is closely related to predicate logic, but has clear advantages wrt. to discourse processing because

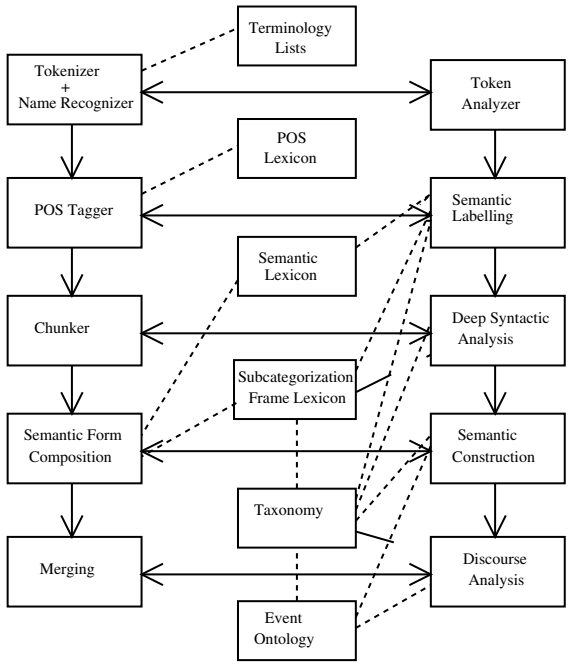

Fig. 2. GenIE's system architecture.

(1) it has proved valuable for discourse representation and the analysis of discourse phenomena such as pronoun resolution [7], presupposition projection [8] and bridging [9], just to name a few.
(2) in contrast to the traditional template representations used in IE, DRT comes with a well-defined model-theoretic semantics [7] for its DRSs.
(3) there exists a sound and complete calculus for first-order DRSs [10] which can be used to define an inference mechanism on DRSs.

Before concentrating on the discourse analysis component of GenIE and the ontology developed for it we will in the remainder of this section shortly describe the other components depicted in Fig. 2.

2.1. Tokenizing and POS tagging

The main task of a tokenizer is to identify and mark sentential boundaries in an input string and to segment these sentences into sequences of tokens. Period disambiguation and thus marking of sentential boundaries is solved to a satisfying extent. We use the tokenizer of Schmid [11], which has an accuracy of about 99.5%. The determination of token boundaries is more complex, however, because—especially in the biochemical domain—technical terms may include a lot of special signs (digits, brackets, periods, etc.). To achieve a similarly high precision a comprehensive lexicon of terminological names and/or a Token-Analyzer are needed.

A lexicon of gene names for *Saccharomyces cerevisiae* and some other eukaryotic organisms is available from http://www.bork.embl-heidelberg.de/synonyms/. These lists have been built up by gathering gene names from various databases (e.g. [12]). The synonym list for *S. cerevisiae* contains 52,059 entries. Generating possible orthographic variants (see Section 2.2 for details), the final list contains 282,323 entries. Since multi-word terms are part of this list, we use them already within the tokenizing step for determining term boundaries. The following levels of processing also benefit from this resource (e.g. the step of semantic labeling, see Section 2.2).

In case a particular token is not found in the lexicon of terminological names, GenIE will deepen its analysis at this particular level. The Token-Analyzer will first of all split the unknown expression into a sequence of morphemes which will then be analyzed as if it were a sentence representation, i.e. the sequence of morphemes will undergo an assignment of sub-lexical POS-tags as well as a sub-lexical syntactic-semantic analysis. As the output of the Token-Analyzer will more often than not be needed for discourse analysis steps at the sentential (or supra-lexical) level a sub-lexical kind of discourse analysis might also be necessary. Such an analysis component should be able to find out that, e.g. 1,3-*Bis*-[3-(*Amino-imino-methyl*)-*phenyl*]-*urea* and 3,3′-*Ureylen-dibenzamidin*, both represent (in a systematic way) the same substance and hence establish coreference between these names (see [13,14]).

Part-of-speech (POS) categories allow to generalize over the syntactic distribution of classes of words. The syntactic analysis profits from the generalization by referring to POS categories instead of referring to specific lexical words. The annotation of words with part-of-speech (POS) categories is strongly interrelated with tokenization, named entity recognition and with syntactic annotation.

In GenIE, we use the TreeTagger developed by Helmut Schmid at IMS Stuttgart [15]. In general POS-tagging deals with two main difficulties, these are (i) disambiguation of ambiguous words [5] and in the absence of a deep analysis of a particular token (ii) guessing of the POS-tags for unknown and unanalyzable words [6]. We conducted an experiment where TreeTagger was trained on a manually annotated training corpus (600 SWISS-PROT functions slots). We evaluated both, the disambiguation accuracy and the guessing performance. For the second task, we trained one version of the tagger with a reduced lexicon by omitting words in the lexicon that exclusively occurred in the testing subcorpus. The tagger was evaluated on unseen testing data (400 function slots, or 9687 token). For the disambiguation task, the accuracy was 98.03%. The inclusion of the guessing task resulted in 94.18% accuracy. Whereby 78.01% of the mismatches were guessing errors. 71.92% of all unknown token were tagged correctly.

2.2. Semantic labeling

After the message understanding conferences (MUC), which were carried out between 1987 and 1997, named entity recognition (NER) is one of the tasks that is regarded as solved to a satisfying extent. [7] Designers of IE systems who have started to attack the domain of biochemistry might have expected similar results. The scientific publications in this area are, however, coined by a huge amount of complex nominal phrases (mainly technical terminology) which in addition have to be distinguished by a substantially bigger number of semantic types. Hence terms that are not listed in the semantic lexicon will have to be analyzed by the Token-Analyzer. The following tasks are of crucial importance for the problem of assigning semantic labels.

(1) The meaning of a term, or its semantic label must be computed on the basis of a morphological and sub-lexical syntactic analysis (as already mentioned above) which is combined with an ontological classification. This analysis supports the identification of the actual entity the term refers to in case the term is not listed in the corresponding database. [8]
(2) Not only for concerns of recall and precision but also for concerns of robustness and performance of the overall system, the task of recognizing syntactic, orthographic and paragrammatical variants of technical terms is of crucial importance in the domain of biochemistry.
(3) Technical terms are highly ambiguous (e.g. protein, gene, enzyme,...). We identified homonyms within bio-related terms, e.g. AGS1 which can either be *activator of heterotrimeric G-protein signaling*, *Aicardi-Goutieres Syndrome gene* 1, or *a yeast gene involved in aminoglycoside antibiotic sensitivity/resistance*. Ags1p stands for *alpha-glucan synthase* 1 *protein*. Homonym problems might not only occur wrt. gene names, but sometimes also between gene names and common english, e.g. *disco*, *boss*, etc.

[5] These are words, which have several POS-tags for one word in the lexicon, e.g. *Map* might either be a noun, a verb or an abbreviation for *mitogen-activated protein*.

[6] Such items get assigned a POS-tag depending on the left POS-context. The underlying probabilistic model is learned by the tagger during a training phase.

[7] *F*-score of more that 90% was quite common for the IE systems. This is still true for typical noun phrases like names of persons or companies, addresses, telephone numbers and the like.

[8] There are databases for chemical compounds, genes, proteins, enzymes and such forth. For each entity they describe a systematic name as well as a set of synonyms is given. Accession numbers provide a unique reference.

Interestingly, we identified in 9000 Medline abstracts 50 different linguistic variations (due to possible combinations of the types of variations in (1) and (2)) of the term *yeast*, or *Saccharomyces cerevisiae*. Attempts for solving these problems mentioned in (3) are currently being developed (see for instance [16–18]).

2.3. Chunking

Text chunking consists of dividing a text in syntactically correlated nonoverlapping, nonrecursive phrases (sentence parts). It is robust and approximates full parsing. In addition precision is at a high level too. [9] In GenIE we make use Steven Abney's partial parser Cass [19]. It is an already approved system that allows the user to modify the cascading set of rules and adjust the system to fit the particular needs of the given text class or domain as we have done for the domain of biochemistry (compare [20]).

Two possible chunk analyses for the phrase *the putative gene for Saccharomyces cerevisiae riboflavin synthase beta chain* are shown in Fig. 3. They both consist of a set of trees, one for the noun phrase *the putative gene* and one for the prepositional phrase *for Saccharomyces cerevisiae riboflavin synthase beta chain*. Note that the trees for the prepositional phrase are not attached to the tree of the noun phrase, because this would lead to recursive structures, which are generally excluded in chunking. The difference between the two trees for the prepositional phrase is that the first one is generated on the basis of lexical entries for each word token, whereas the second tree results from a lexicon that contains *Saccharomyces cerevisiae* and *riboflavin synthase beta chain* as multi-word tokens with semantically labeled POS tags nn_{org} and nn_{prot}.

As we cannot assume that the lexicon of multi-word terms is complete it should in principle be possible to reconstruct the additional structure, or grouping, by a deeper syntactic and semantic analysis. Similarly to the interpretation strategy of the Token-Analyzer this deeper analysis will parse the full prepositional phrase with the result shown in Fig. 4. This deeper syntactic and semantic analysis component is described in the next section.

2.4. Deep syntactic analysis and semantic construction

The aim of the deep syntactic and semantic analysis is on the one hand to yield a full parse for the sentence in question and on the other hand to build up a semantic representation of it. In this sense it is also used to connect together the partial syntactic structures produced by the chunker to yield a complete syntax tree as in Fig. 4. Thus, a semantic representation can also be constructed out of the output of the chunker.

As already mentioned, our semantic construction component builds a DRT representation for each sentence in the text. For this purpose, it makes use of an LTAG-based approach [21] as described in [22]. The key of the approach is that every lexical element is associated with a so called *elementary tree* containing basic semantic and ontological information about it as well as information about how it can be combined with other elements. This information is stored in the semantic lexicon (compare Fig. 2) and—to a great extent—obtained from the sub-categorization

[9] Precision of more than 90% are the rule.

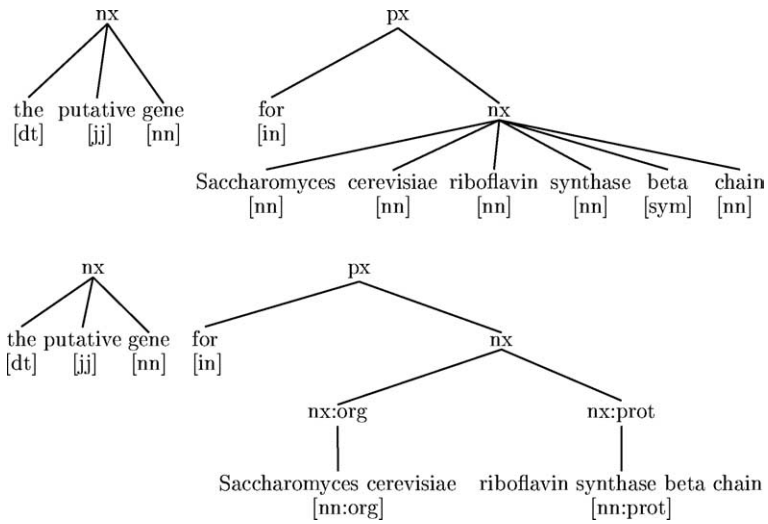

Fig. 3. Two possible chunk analyses for the putative gene for *Saccharomyces cerevisiae* riboflavin synthase beta chain.

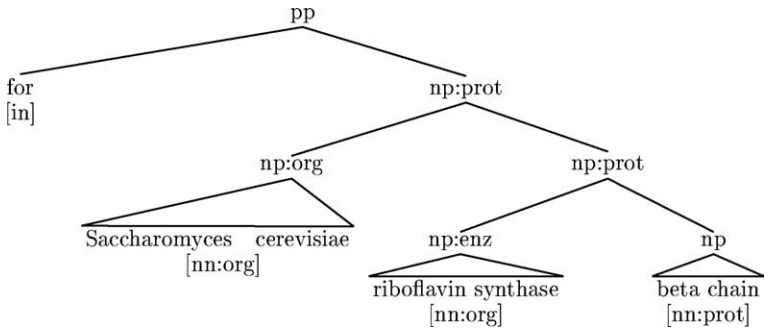

Fig. 4. Result of the deeper syntactic analysis.

acquisition component (see below). It is important to mention that each lexical element can be associated with a set of different elementary trees. In fact, an important task of our semantic construction algorithm is to choose the correct elementary tree for each word in the processed sentence. Thus, our deep syntactic and semantic analysis also performs lexical disambiguation. Fig. 5 shows the relevant elementary trees for the words in the first sentence of Example (1).

In the figure every node is marked with a label according to its syntactic category. Furthermore, if appropriate, a node is marked with its corresponding ontological category following the ':' symbol. The $\leqslant_C$ sign is used to denote that any node identified with the node in question needs to be from an ontological point of view identical or more special than it (compare [22]). The semantics of each node can be found below it. Note that we reify events as in [23] and introduce a discourse referent—e in the figure—for them. As already mentioned, the semantic representation language used is DRT and a pair $\langle D,C\rangle$ represents a DRS where D is the set of discourse referents and C is the set of conditions on them (compare [7]). In particular, in our approach we make use of an

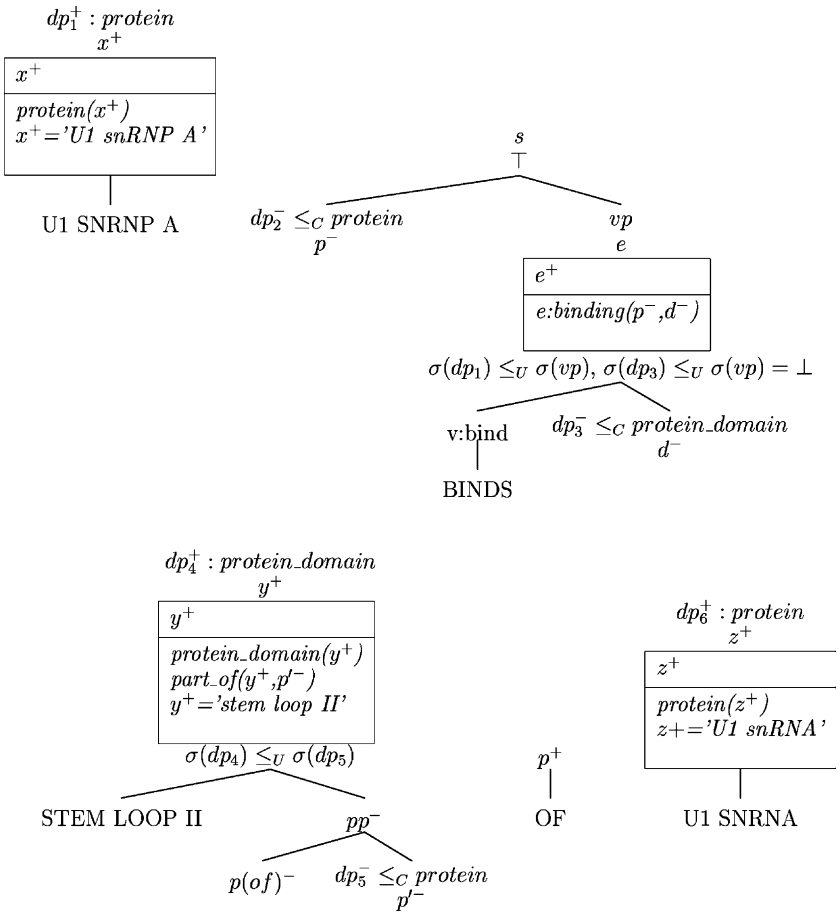

Fig. 5. Elementary descriptions for U1 SNRNP A BINDS STEM LOOP II OF U1 SNRNA.

extension of DRT called UDRT (Underspecified Discourse Representation Theory). UDRT [24] introduces a partial order $\leqslant_U$ between DRSs to explicitly talk about scope relations. This partial order can in some cases be strengthened to a linear one in order to yield a (standard) DRS. In line with [25] nodes are marked with a '+' or '−' sign. Nodes with a '+' sign are anchored to some lexical element, while those with a '−' sign are not. The nodes without a '+' or '−' sign are said to be saturated, i.e. positively and negatively anchored. As in [25], every node has to be lexically anchored and thus positively and negatively marked nodes have to be identified. The root (s in our case) is supposed to be positively and negatively marked. In the extension we proposed in [26], we additionally talk about discourse referents which can also be positively and negatively marked. By analogy to [25] they also need to saturated, i.e. positively and negatively marked. The verb BINDS for example introduces one saturated event discourse referent 'e' but also two negatively marked discourse referents representing the arguments to be filled. Further, we functionally assign discourse referents to nodes such that if the corresponding nodes are identified, their discourse referent will be identified as well. The semantic representation—a DRS—is also functionally assigned to a node via the function σ. In Fig. 5 the corresponding discourse referents

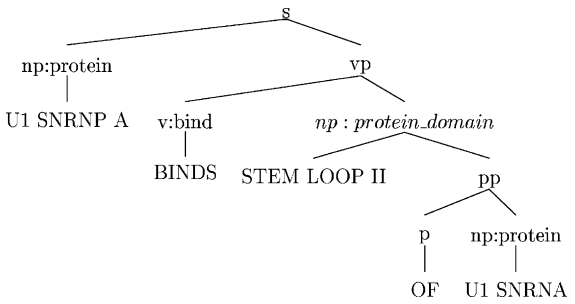

Fig. 6. Syntax tree for U1 SNRNP A BINDS STEM LOOP II OF U1 SNRNA.

and DRSs are given below each node. Note that certain dominance relations in terms of the partial order $\leqslant_U$ are also given under the nodes. The reader is referred to [26] for details which are out of the scope of this paper.

By identifying nodes in this way we get the syntax tree in Fig. 6 for the first sentence in Example (1) as well as the following DRS by strengthening the partial order to a linear one:

e,p,d,p'
e:*binding*(p,d)
protein(p)
p = '*U1 snRNP A*'
protein_domain(d)
d = '*stem loop II*'
part_of(d,p')
protein(p')
p' = '*U1 snRNA*'

An important question is how to deal with implicit arguments of a certain verb. As shown in [27] this is not a trivial task. For this purpose we do actually not require the mapping between positively and negatively anchored nodes to be 1–1 as in [22] and map the verb to the elementary tree with the most subcategorized arguments such that the ones which are not positively anchored to lexical elements can be 'filled' later in the discourse. On the basis of the output of this component, the discourse analysis component computes conceptual relations between the sentence DRSs (compare Section 5).

2.5. Subcategorization frame acquisition

Central to our approach is the availability of *sub-categorization frames* for each verb. A subcategorization frame specifies for a given verb the number, the syntactic/semantic category and the thematic role of each of its arguments. It is important to mention in this context that a verb can have different sub-categorization frames, possibly also with different meanings. The sub-categorization frame for the verb *bind* in the first sentence of Example (1) could be represented as follows:

bind(*subj*(*agent*):*protein*,*obj*(*patient*):*protein domain*). The specification of the semantic/ontological class of the arguments of a certain verb is widely known as its *selectional restrictions*.

An important question in this context is certainly how to acquire the possible sub-categorization frames for all the relevant verbs. One possibility would certainly be to handcraft them by manually analyzing the corpus, but this would obviously be a too time-consuming and tedious task. Thus, in our approach these sub-categorization frames are automatically acquired from a domain-specific corpus. For this purpose, we first parse the corpus with LoPar [28], a statistical left-corner parser, and extract for each sentence the main verb together with the head of the arguments it subcategorizes with *tgrep*. [10] Then, we compute for every different argument structure of each verb the number of times that a certain concept in the ontology appears at a certain argument position, i.e. as object, subject, etc. For each different argument structure we then build the corresponding sub-categorization frame by selecting the concept which appears with the highest frequency at each argument position. It is important to mention that we do not only sum up over the explicit occurrences in the corpus, but also take into account the concept hierarchy as in [29] in order to yield sub-categorization frames with the most general selectional restrictions (compare [30]). For this generalization purpose we make use of a not publicly available concept hierarchy for the biochemical domain developed at the Europan Media Laboratory. [11]

3. A corpus study

Swiss-Prot is an annotated protein sequence database. It is composed of sequence entries which in turn are composed of different line types each with their own format. The DE (DEscription) line for example contains general descriptive information about the sequence. In particular it gives the proposed official name as well as synonyms for the protein sequence in question. On the other hand, the CC line contains free text comments on the entry. It is further divided into different topics. The CC FUNCTION topic for example consists of natural language descriptions of the protein's function.

A corpus has been built containing the DE line and the CC FUNCTION topic of 20,189 Swiss-Prot database entries. In the following, this corpus will be referred to as 'the Swiss-Prot corpus'. The length of the *CC FUNCTION*-slot is between 1 and 26 sentences with an average of 1.6. The length in words ranges from 1 to 172 and is 22 on average.

As a first step, we decided to concentrate on the analysis of binding events as expressed by the second most frequent verb *binds* and its gerund *binding* (both together constituting 4.5% of the verbal forms of the corpus) as the meaning of the most frequent verb *involved* (6.2%) is too dependent on what something is involved in and thus it is difficult to decide whether a certain expression can be understood as standing in a conceptual relation to it or not. Furthermore, it is not clear if the verb *involved* has an event reading at all. From the authors' point of view it denotes rather a state than an event (see [7] for a formal definition of states and events).

So all the entries from the Swiss-Prot corpus containing the verbs *bind*, *binds* and *binding* have been selected. Out of the resulting 3623 entries, 500 have been randomly chosen. A detailed study

[10] See http://mccawley.cogsci.uiuc.edu/corpora/treebank3.html.
[11] http://www.eml.org.

Table 1
Results of the classification of binding events as antecedents

Type	Occurrences
Event coreference	27 (5.1%)
Event bridge	137 (25.9%)
Event role	28 (5.3%)
Total binding events	528 (100%)

of these entries allowed to distinguish three relationships between events as antecedents and some other event, state or entity as referring expression:

- As *event coreference* will be regarded the identity relation between a linguistic expression representing an event e_2 and the antecedent event e_1 it refers to, i.e. $e_1 = e_2$, such as in Example (1).
- As *event bridge* [12] will be regarded the nonidentity relation R between a linguistic expression representing an event, state or entity e_2 and some antecedent event e_1, i.e. $R(e_1, e_2)$, as in Example (2).
- The relation between an expression representing an entity e_2 referring to a (possibly implicit) argument of an antecedent event e_1 and the event in question, i.e. $\text{Role}(e_2, e_1)$, will be called *event role*. Here is an example:

(3) TRANSCRIPTIONAL ACTIVATOR THAT BINDS TO THE ENHANCER OF THE ADENOVIRUS E1A GENE; THE CORE-BINDING SEQUENCE IS 5′[AC]GGA[AT]GT-3′.

One of the authors has classified the binding events of the 500 entries mentioned above into the three suggested categories. The results are summarized in Table 1 and show that well above one third of the binding events in the corpus represent an antecedent for some other expression. Thus the necessity of resolving conceptual relations between events, states or entities to events as antecedents in order to establish discourse coherence becomes also clear from a quantitative point of view. In order to verify the utility and scalability of the approach presented within this work, a quantitative measurement of its performance has been carried out. Typically within computational linguistics research and in particular in the field of information extraction, such an evaluation of the performance of an approach involves the development of it on training data and the subsequent verification of its scalability on unseen or test data.

For this purpose, the above mentioned 500 Swiss-Prot entries have been divided into a training and a test corpus each consisting of 250 entries. In both the training and test corpus verbs and definite descriptions (DDs) representing events, states or entities have been marked by one of the authors and assigned a unique identifier. Table 2 gives some statistics about the training

[12] This nomenclature is introduced by analogy to the famous *bridging* phenomenon [31,32].

Table 2
Statistics of the training and test corpora

	Training corpus	Test corpus
#Tokens	12,666	12,180
#Events	708 (54.05%)	894 (56.69%)
#States	175 (13.36%)	209 (13.25%)
#Entities	427 (32.60%)	474 (30.06%)
Total	1310 (100%)	1577 (100%)
#DDs	510	530

and test corpora. In particular it indicates the number of tokens, the number of events, states and entities marked as well as the number of definite descriptions of each corpus.

4. An ontology of biochemical events

An ontology is a specification of a conceptualization [33]. A conceptualization can be understood as an abstract representation of the world or domain we want to model for a certain purpose. From a formal point of view, it will be understood as a triple $O = (C, T, D)$, where C is a set of concepts relevant for the domain in question, T is a set of taxonomic relations defined on the concepts in C and D is a set of partial definitions of concepts in the sense that they specify their necessary conditions [33]. In what follows, we describe the principles under which an ontology of biochemical events for the discourse analysis component has been developed.

4.1. Classifying biochemical verbs

The first step consisted in a rough semantic classification of the biochemical verbs appearing in our training corpus. An intense corpus study allowed to identify 11 different semantic classes of verbs together with their corresponding pre- and post-conditions. These semantic classes are: *control/regulation*, *biochemical interaction*, *logical interaction*, *biochemical process*, *binding/dissociation*, *formation*, *integrity*, *modification*, *availability*, *change of location*, *temporal order*. A detailed description of each semantic class, the list of all the verbs contained in it as well as a formal specification of its pre- and post-conditions can be found in [34]. In this paper we will focus on the *control/regulation* and *biochemical interaction* classes in order to illustrate the performed classification.

From a general point of view it can be asserted that proteins usually control or regulate biochemical processes in the sense that they affect them by stimulating, activating or inhibiting them. However, it seems quite difficult to specify the pre- and post-conditions of such a *control/regulation*-event in general. A control/regulation can actually affect the speed as well as the concentration of the products or other properties of a certain reaction or biochemical process. Thus a *control/regulation*-event can be intuitively formalized as changing the value of some measurable property of a certain biochemical process:

$$\forall e,p,b((e : control/regulate(p,b) \wedge protein(p) \wedge biochemical\_process(b)) \\ \iff \exists P,s,s'(measurable\_property(P) \wedge s : value(P(b)) = v_1 \\ \wedge s \supset\subset e \wedge s' : value(P(b)) = v_2 \wedge v_2 \neq v_1 \wedge Result(e,s'))) \quad (4)$$

The above formula states that e is an event in which the protein p controls/regulates a certain biochemical process b if and only if there is a measurable property P of this biochemical process which has different values in the state s, i.e. before the event takes place, and in the state s', i.e. after the event takes place. In this sense $s \supset\subset e$ means that the state s is temporally before e and that there is no other time interval between them (compare [7]). The notation $s{:}value(P(bp)) = v_1$ is nothing else that another way of writing $value(s,P(bp)) = v_1$. Result(e,s') is the rhetorical relation defined by Lascarides et al. [35] and its meaning is that the event e causes the event or state s' as a result. Furthermore, Result(e,s') also implies $e \supset\subset s'$.

Chemical substances interact with each other in various ways. These interactions normally result in the change of some biochemical property of some of the implicated substances. Thus, a *biochemical interaction* between two substances s_1 and s_2 can be defined as changing some *biochemical property* P of one or both of the substances involved in the interaction:

$$\forall e,s_1,s_2((e : biochemical\_interaction(s_1,s_2) \wedge chemical\_substance(s_1) \wedge chemical\_substance(s_2)) \\ \iff \exists P,s,s'(biochemical\_property(P) \wedge (s : \neg P(s_1) \wedge s' : P(s_1) \\ \vee s : \neg P(s_2) \wedge s' : P(s_2)) \wedge s \supset\subset e \wedge Result(e,s'))) \quad (5)$$

4.2. Building a taxonomy of biochemical events

In this section we explain how the development of a hierarchical taxonomy of events can be systematically grounded upon the specification of the pre- and post-conditions of semantic classes as described above. First, we will argue that according to our axiomatization of a *control/regulation* event as given above, an *inhibition* can be seen as more special than it. In fact an *inhibition* event can be seen as reducing some measurable property related to the effects of some biochemical process, i.e.

$$\forall e,p,b((e : inhibition(p,b) \wedge protein(p) \wedge biochemical\_process(b)) \\ \iff \exists P,s,s'(measurable\_property(P) \wedge s : value(P(b)) = v_1 \\ \wedge s \supset\subset e \wedge s' : value(P(b)) = v_2 \wedge v_2 < v_1 \wedge Result(e,s'))) \quad (6)$$

Thus the pre- and post-conditions of an *inhibition* event are more special that the ones of a *control/regulation* event, such that we can add the following axiom to the set of taxonomic relations T:

$$\forall e,p,b((e : inhibition(p,b) \wedge protein(p) \wedge biochemical\_process(b)) \\ \rightarrow e : control/regulation(p,b)) \quad (7)$$

As a second example we also show that the binding of two proteins can be seen as some sort of *biochemical interaction* between them. This is indeed the case if we regard $\lambda p_1.bound(p_1,p_2)$, i.e. the property of being bound to some protein p_2 as a biochemical property, i.e.

$$\forall p_1(protein(p_1) \rightarrow biochemical\_property(\lambda x.bound(x, p_1))) \tag{8}$$

and thus as specializations of the relation P in definition (5). As a result, the following taxonomic relation can be added to the set of taxonomic relations T:

$$\forall e, p_1, p_2((e : binding(p_1, p_2) \wedge protein(p_1) \wedge protein(p_2)) \rightarrow e : biochemical\_interaction(p_1, p_2)) \tag{9}$$

Of course, the same argumentation as above then holds for a *dissociation* event so that the whole *binding/dissociation* class can in fact be seen as a specialization of the *biochemical_interaction* class. So it is obvious that the classes proposed above are not disjoint. The reason is that the classes are linguistically motivated in the sense that the aim of the classification has been to group as much verbs as possible into a reasonable small number of classes and then give an appropriate and general semantics for each of these classes. If the aim had been to yield a disjoint classification of events, the proceeding would have had to be the other way round, i.e. first a number of conceptual and disjoint classes would have had to be identified and then suitable verbs would have had to be found for each class.

4.3. Conceptual definitions

Before getting into details concerning the principles and methods used to model the relevant conceptual definitions of the domain in question, it should be first explained what a conceptual definition is supposed to be. The view of conceptual definitions underlying the work presented here is that they describe the nature of a certain concept or object by specifying other concepts or objects which are related to it in a specific way. When applying this idea to the domain of molecular biology we could say that it is in the nature of a binding between two proteins to produce a complex as a result. Similarly, from a general point of view, it could be claimed that it is in the nature of a protein interaction to regulate some other biochemical processes as a result.

This fact can be represented by the following axiom in the set of conceptual definitions D:

$$\forall e, p_1, p_2((e : biochemical\_interaction(p_1, p_2) \wedge protein(p_1) \wedge protein(p_2)) \rightarrow \exists b, e'(e' : control/regulation(p_1, b) \wedge biochemical\_process(b) \wedge Result(e, e'))) \tag{10}$$

The above formula can be paraphrased as follows: for each *biochemical interaction* between two proteins, there is some event in which one of the involved proteins controls/regulates some other biochemical process as a result of their interaction.

4.4. The ontology O_{Bio}

After having described the methodology according to which an ontology of events can be developed, we now describe the concrete ontology $O_{\mathrm{Bio}} = (C_{\mathrm{Bio}}, T_{\mathrm{Bio}}, D_{\mathrm{Bio}})$ which was developed on the textual basis of the training corpus in the sense that suitable conceptual DRT-representations of events, states and entities appearing in the corpus have been developed by the authors. These conceptual representations constitute the set C_{Bio}. Furthermore, following the principles described in the previous section, the set T_{Bio} specifying the taxonomic relations between these concepts has been constructed and conceptual relations between different events or states have been captured

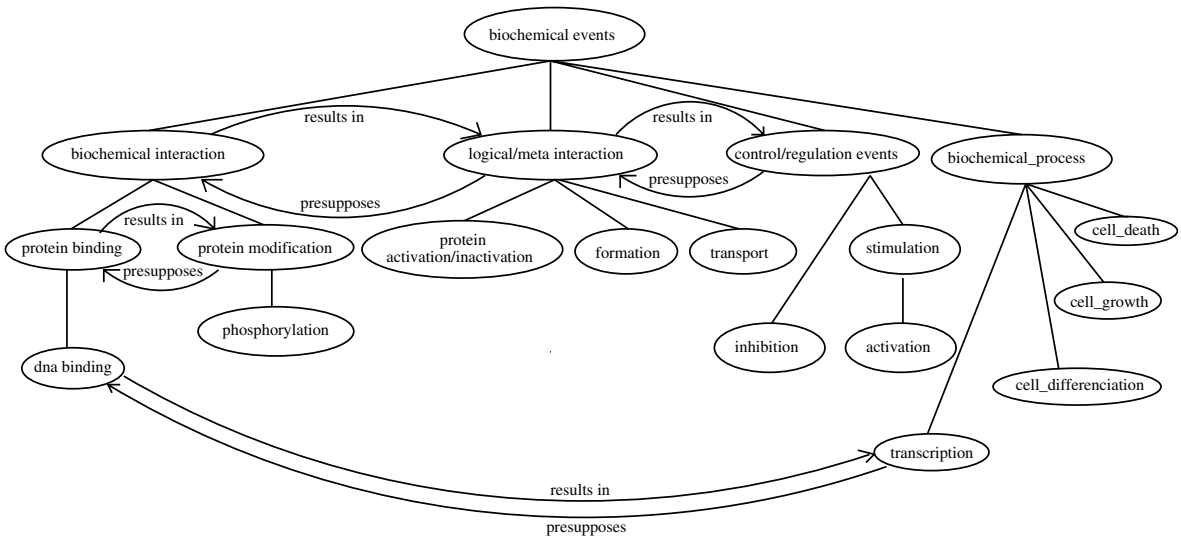

Fig. 7. Top level of the ontology O_{Bio}.

in form of logical axioms. These axioms form the set D_{Bio} consisting of partial definitions of concepts.

Within this ontology development step, special attention has been paid to represent only those concepts as well as those taxonomic and conceptual relations having a certain degree of relevance and generality. The aim has been to yield an ontology which is not too specifically tailored to the corpus it was developed on thus being potentially reusable across different biochemical texts. The ontology developed has 129 concepts ($|C_{\text{Bio}}|$), 50 taxonomic relations ($|T_{\text{Bio}}|$) and 19 axiomatic definitions of concepts ($|D_{\text{Bio}}|$). Fig. 7 shows graphically the top concepts, the basic taxonomy as well as some of the relations in the ontology. [13] After developing the ontology, the marked events, states and entities of both the training and test corpus have been manually mapped by the authors to DRSs representing the corresponding ontological concept in C_{Bio}.

It is important to mention that this small ontology was created merely in order to test the ontology-driven approach to discourse analysis presented in this paper and in order to verify its potential usability. For this reason, we will gloss over the details concerning the developed ontology. The development of a suitable and broad coverage ontology for the domain of molecular biology is definitely out of the scope of the work presented here. The interested reader is referred to Ratsch et al. [36].

5. The ontology-driven approach to discourse analysis

The ontology-driven approach presented in this paper makes use of a semantic representation of the text such as provided by the semantic construction component described in Section 2 to

[13] The complete ontology can be found in [34].

make contextual information explicit as well as of a model of the domain in form of an ontology as described in the previous section to infer conceptual relations between events as antecedents and other events, states or entities as referring expressions. In principle, the idea behind the approach is that the ontology specifies the way how the events extracted from a text are conceptually related to each other.

The approach presented in this paper is implemented in Prolog as a variant of the presupposition projection algorithm described in [37]. In line with [8] and [9], referring or anaphoric expressions are represented by α-marked DRSs [14] which have to be linked to a previous suitable antecedent and thus are resolved. The approach presented here differs from Bos et al.'s in the sense that it does not only consider definite descriptions as presupposition triggers, i.e. as referring expressions, but also verbs representing events and states in the sense that they are normally related to a previous event thus establishing discourse coherence [32]. The most important difference, however, is that it makes use of an ontology of events replacing Bos et al.'s *qualia structure* [9]. In contrast to the qualia structure, the ontology does not only represent lexical knowledge, but complex world knowledge about events.

On the basis of an ontology as defined in the previous section, [38] defines a notion of specialization $<_O$ between DRSs. The concepts in C are represented as DRT-based predicate argument structures (compare Section 2).

Here are the definitions for the resolution of the three different relations considered in Section 3:

Definition 1 (*Event coreference*). Two events e_1 and e_2 appearing in the text (in this order) and respectively represented by the DRSs K_1 and K_2 will be linked by Coreference, i.e. $e_1 = e_2$, iff K'_1 is an ontological generalization of K_1, i.e. $K_1 \leqslant^*_O K'_1$, and furthermore K_2 is suitable to K'_1, where $\leqslant^*_O$ is the reflexive and transitive closure of $<_O$ and *suitability* defines a homomorphism [15] on DRSs as in [38].

The definition captures the intuition that certain expressions are referred to in a more general way later in the discourse, such as in Example (1).

Definition 2 (*Event bridge*). An event e_1 and an eventuality, i.e. an event or state e_2 appearing in the text (in this order) and respectively represented by the DRSs K_1 and K_2 will be linked by the relation R, i.e. $R(e_1,e_2)$, iff K'_2 is an ontological generalization of K_2 and K'_1 follows logically from K_1 with regard to the ontology as defined in [38], i.e. $K_2 \leqslant^*_O K'_2$ and $K_1 \Rightarrow_O K'_1$ and K'_2 is suitable to K'_1, where $\oplus$ is the merging operator for DRSs [7].

With the above definition and if a binding (of a protein to DNA) is regarded as a *biochemical interaction* and furthermore a *biochemical interaction* is defined as leading to a *control regulation* and an *inhibition* is more special than it (compare Section 4), then Example (2) can be successfully resolved.

Definition 3 (*Event role*). An event e_1 and an entity e_2 appearing in the text (in this order) and respectively represented by the DRSs K_1 and K_2 will be linked by the relation $\mathrm{Role}(e_2, e_1)$ iff K_2

[14] These are DRSs marked as unresolved, i.e. which have to be resolved with regard to the preceding context [9].

[15] Homomorphism is understood here in a mathematical sense, i.e. as a group preserving operation on DRSs. It can be actually seen as a matching operation between DRSs.

matches a (real) subset of the conditions of K_1, i.e. K_2 is suitable to K_1', where $\mathrm{Con}(K_1') \subset \mathrm{Con}(K_1)$ and $\mathrm{Con}(K)$ are the conditions of the DRS K as defined in [7].

The above definition obviously presupposes that the implicit roles of each event are made explicit in the representation of the corresponding concept. Assuming for example that the binding sequence is modeled as an implicit role, i.e. an attribute of a DNA binding event in the ontology, Example (3) can be successfully resolved. In general, in the approach presented here, reference resolution is made determinate by choosing the most recent antecedent and minimizing reasoning complexity with regard to the ontology [38]. In addition, we will prefer resolutions as *coreference* over *bridging event* and *event role*.

6. Evaluation

6.1. The task

The task on which the ontology-driven approach presented in this paper has been evaluated can be stated as follows: given a short text from Swiss-Prot describing the function of a protein as well as an ordered list of DRSs representing events or states defined with regard to an ontology O_{Bio} and assumed as already extracted from this text and thus representing its discourse structure, can we infer the correct conceptual relations between these events or states? The conceptual relations considered are the *event role*-relation between an entity and an event, *event coreference* between two events as well as the following two instances of the generic *event bridge* relation: *Result* [39] and *Explanation/Elaboration*, where the latter is defined as the disjunction of the Explanation and Elaboration relations considered by Lascarides et al. [39]. The reason why they have been collapsed into one relation is that the distinction between them has been expected to be difficult for the annotators.

6.2. Agreement between annotators

In order to evaluate the performance of the discourse analysis component in a quantitative manner, the training and test corpora have been annotated by different subjects with the above introduced discourse relations by making use of the MMAX annotation tool developed by Müller et al. [40]. The relevant events, states and entities had been previously marked by the authors so that the task of the annotators has basically been to choose the appropriate conceptual relation between two marked expressions.

The training corpus has been annotated only by one of the authors (AUTHOR in the following), while the test corpus has been annotated independently from each other by the AUTHOR and two biologists. The agreement between the annotators has been measured with the kappa statistic [41]. The overall kappa coefficient has been determined to $K = 0.31$. Following the classification by Landis and Koch [42] of the agreement as measured by the kappa statistic, this value can be classified as corresponding to a 'fair' agreement between the annotators. Certainly, the agreement is not good enough for tentative conclusions to be drawn [41], which is *per se* an interesting result. It furthermore hints at the fact that the experiment should be reconsidered and

redone with a modified and probably simpler version of the proposed classification task. On the other hand, the low agreement shows that the task of determining discourse relations specifying the way how discourse segments are connected together is not a trivial one and that it is quite subjective. This observation already points to the limits of a machine-based approach.

6.3. Results

The performance of the approach outlined in Section 5 on the training and test corpus has been measured in terms of precision and recall against a certain standard. The recall (R) is a measure of how many of all the possible correct answers are found by the approach, while the precision (P) is a measure of how many of the total answers given are actually correct. The *F-measure* is a metric which combines recall and precision into a single value.

$$R = \frac{\#\text{ correct answers}}{\#\text{ possible answers}} \quad P = \frac{\#\text{ correct answers}}{\#\text{ given answers}} \quad F = \frac{2 * P * R}{P + R}$$

The approach described in Section 5 yielded a recall of $R = 52.57\%$ and a precision of $P = 84.40\%$ and thus $F = 64.79\%$ measured against the AUTHOR's annotation of the training corpus. The performance of the approach on the test corpus has been measured against the following four standards:

- AUTHOR: the set of discourse relations annotated by the AUTHOR,
- 2/3: the set of discourse relations on which at least two of the three annotators agree,
- 3/3: the intersection of the discourse relations of all the annotators, i.e. the ones on which all three agree,
- UNION: the union of the discourse relations of all three annotators.

Table 3 indicates the recall and the precision measured on the four test standards defined above. The recall on the AUTHOR, 2/3 and 3/3 standards seems quite reasonable ranging from 45.38% to 54.54%. It is interesting to observe that the highest recall of 54.54% corresponds to the standard containing those relations annotated by all the three subjects, so that it can be concluded that the system is in fact computing most of the relations that all annotators agree on, i.e. the most reliable ones. The precision values are actually much worse. This is without doubt due to the low agreement of the annotators as the system is actually computing relations which have been annotated by only one of the annotators and therefore neither appear in the 2/3 nor in the 3/3 standard. Thus the system is being penalized for finding relations which have been annotated by some annotator

Table 3
Results of the bridging reference resolution approach measured against the four standards: AUTHOR, 2/3, 3/3 and UNION

Standard	Cardinality	Recall (%)	Precision (%)	*F*-measure (%)
AUTHOR	184	53.84	79.84	64.29
2/3	154	45.38	47.58	46.45
3/3	33	54.54	14.52	22.93
UNION	676	16.54	90.32	27.96

Table 4
Conjunctions and the corresponding discourse relations inferred from them

Conjunction	Discourse relation
after	Narration
also	Narration
because	Explanation/Elaboration
but	Contrast
by	Explanation/Elaboration
then	Narration
thereby	Result
therefore	Result
through	Explanation/Elaboration
thus	Result
via	Explanation/Elaboration

and could actually be correct. These observations lead us to also consider the union of the relations annotated by all of the subjects. The precision on the UNION standard was actually quite good (90.32%) such that it can be concluded on the one hand that the major bottleneck of the experiment is in fact the bad agreement between annotators. But on the other hand it nevertheless has to be concluded that the system is performing reasonably well, i.e. getting well above 50% of the most reliable relations and computing less than 10% relations which actually have to be regarded as incorrect.

6.4. Exploiting lexical clues in the resolution process

A further interesting observation is that in many cases there are lexical clues which already indicate the conceptual relation between two eventualities. This is in particular the case for conjunctions such as *by*, *thus*, *because*, *also*, just to name a few. Take for instance the following example:

(11) ALPHA-CONOTOXINS ACT ON POSTSYNAPTIC MEMBRANES, THEY BIND TO THE NICOTINIC ACETYLCHOLINE RECEPTORS (NACHR) AND THUS INHIBIT THEM.

This observation has lead to the idea that discourse relations could also be lexically inferred. [16] For this purpose, the semantic representation of the text has been enriched with a predicate specifying the lexical element by which events are connected. On the basis of such a representation rules have been defined for example stating that if two events are lexically connected via the conjunction *thus*, then normally *Result* is the relation between them.

In this sense, a lexicon containing conjunctions as well as the corresponding discourse relation which can be 'lexically' and nonmonotonically inferred from them has been built (see Table 4). The *Narration* and *Contrast* relations [35] can not be inferred by the ontology-driven approach. However, they are considered to rule out inconsistencies. Then the test corpus has been annotated

[16] Obviously this does not work for all relations, in particular not for *Identity*/*Coreference* and *Role*.

Table 5
Results of the combination of the ontology-driven and the lexically driven approach

Standard	Recall (%)	Precision (%)	*F*-measure (%)
AUTHOR	61.41	76.87	68.28
2/3	48.70	51.02	49.83
3/3	63.63	14.29	23.34
UNION	20.24	93.19	33.26

with a special *connect*-predicate making the conjunction specified in this lexicon between two events explicit. Furthermore, a simple approach has been developed which, given a specific instance of a *connect*-predicate specifying the conjunction linking two events together, nonmonotonically infers the corresponding discourse relation from the lexicon. This 'lexically driven approach', as it will be referred to, yielded a very high precision measured against the UNION standard (100%) but very low recall values measured on the other three standards (16.23%–18.18%).

These results suggest that most of the discourse relations in the corpus can not be inferred by lexical means and show that a knowledge-based approach is in fact necessary. Nevertheless the results also suggest that the ontology and lexically driven approaches could be combined somehow to increase the performance of the whole discourse analysis component. Thus, the decisive question is how to combine the set A of 'ontologically inferred' and the set B of 'lexically inferred' discourse relations. In fact, taking into account the low recall of the lexically driven approach, it seems obvious that the set A will basically determine the overall recall of the system while B will be responsible for increasing the overall precision by eliminating incorrect relations from A. The formula by which both approaches have been combined is $C = A \cup B - \text{inconsistent}(A, B)$, where $\text{inconsistent}(A, B)$ is the set of elements of A and B which given a certain referring expression differ in the corresponding conceptual relation between this expression and some antecedent. Table 5 presents the results of the combination of both approaches and clearly shows that it increases not only the precision but also the recall of the whole approach. The recall for example ranges on the AUTHOR, 2/3 and 3/3 standards from 48.70% to 63.63% and is thus higher when compared to the purely ontology-driven approach. When considering the precision on the UNION standard it can be stated that is has definitely increased. In terms of the arguments given in the previous section it can be asserted that the system is computing almost two thirds (63.63%) of the most reliable discourse relations and that in only less than 7% of the cases the computed relations have to be regarded as actually incorrect. The conclusion is that the lexically driven approach outlined in this section can in fact complement the ontology-driven approach and definitely improve the overall performance of the system.

7. Discussion and related work

It could be certainly argued that this ontology-driven approach to discourse analysis making use of semantic discourse representation structures to represent the linguistic context is not within the scope of the information extraction task as envisioned by Appelt et al. [6]. However, recent work in IE [43,44,4] has shown that in certain domains the whole text is relevant so that

the difference between information extraction and text understanding seems not that relevant anymore. This is also the view underlying this work. On the other hand, Huttunen et al. [44] clearly motivate a discourse analysis as proposed in this paper. They report that in their *Natural Disaster* and *Infectious Disease Outbreak* scenarios the relevant facts are scattered through the whole texts and also express the need to identify relations of inclusion or causation between these facts [44]. This is exactly the aim of the approach presented in this paper. However, the approach is not restricted to the computation of inclusion or causation relations, but to any relation defined within a given ontology. Furthermore, a lot of work is being done concerning the development of suitable ontologies for the domain of biology [45–47,36] so that detailed and broad coverage ontologies to be exploited within such an approach can be expected to be available in the near future.

Discourse analysis within information extraction systems typically boils down to entity coreference resolution and template merging as defined in the MUC tasks. Humphreys et al. [48] and Yangarber et al. [49] present a knowledge-based approach to coreference resolution making use of an explicit semantic representation in form of a predicate-argument structure as well as a taxonomy of concepts. The results of the LaSIE system [48] on the entity coreference task were a recall of $R = 50.71\%$ and a precision of $P = 71.93\%$ on the MUC-6 management succession scenario and $R = 56.1\%$ and $P = 68.8\%$ on the MUC-7 launch event scenario. The results of the Proteus system on the entity coreference task of the MUC-6 management succession scenario were a recall of $R = 53\%$ and a precision of $P = 62\%$ [49].

The above results are not directly comparable to the ones of the approach presented in this paper due to several reasons. First, the domain of application is different from the one of all the other systems. Second, most systems concentrate on the resolution of coreferences between objects or events but none of them attempts to compute discourse relations between events, so that the task at hand seems inherently harder. Third, the approach presented here has been evaluated given a semantic representation of the text, while the other systems have been evaluated either given a syntactic representation or even raw text. Nevertheless, a comparison between the results of the approach presented here and the ones discussed shows that from a quantitative point of view it fits quite well in the picture of the results of other systems dealing with a similar task in the field of discourse analysis in IE.

8. Conclusion and further work

This paper has presented an ontology-driven approach which on the basis of a given ontology as well as a semantic representation of the events extracted from a text, computes conceptual relations between these events and a referring expression representing some other event, a state or an entity. It has furthermore outlined a lexically-based approach which can actually complement the ontology-driven approach improving its results in terms of recall and precision. The overall results of the approach are very promising and are comparable to other systems dealing with discourse phenomena such as coreference resolution. It is important to mention that the approach presented here is not inherently restricted to DRT as discourse representation language. As long as an inference mechanism and a notion of homomorphism and accessibility can be defined with regard to

some other semantic representation structures, they can definitely replace DRT. On the other hand it would also be interesting to explore more refined inference mechanisms as well as to address the problem of acquiring ontological relations automatically from text or other sources.

Acknowledgement

We would like to thank anonymous reviewers of the NLDB 2003 conference as well as of the DKE Journal for comments on the paper We would also like to thank Anette Kurz and Friederike Niestroj for annotating the Swiss-Prot function descriptions as well as Esther Ratsch and Ulrike Wittig for answering now and then our biochemical questions. Furthermore, we are grateful to Michael Strube for advice concerning the evaluation of the system as well as to Christoph Müller for help with the annotation tool MMAX. Finally, Philipp Cimiano would like to acknowledge the European Media Lab and in particular Isabel Rojas for kindly allowing him to use its facilities.

References

[1] T.C. Rindflesch, J.V. Rajan, L. Hunter, Extracting molecular binding relationships from biomedical text, in: Proceedings of the ANLP-NAACL 2000, 2000, pp. 188–195.
[2] J. Pustejovsky, J. Castaño, J. Zhang, Robust relational parsing over biomedical literature: Extracting inhibit relations, in: Proceedings of the Pacific Symposium on Biocomputing (PSB'02), 2002, pp. 362–373.
[3] A. Yakushiji, Y. Tateisi, Y. Miyao, Event extraction from biomedical papers using a full parser, in: Proceedings of the Pacific Symposium on Biocomputing (PSB'01), 2001, pp. 408–419.
[4] U. Reyle, J. Saric, Ontology driven information extraction, in: Proceedings of the 19th Twente Workshop on Language Technology, University of Twente, 2001, pp. 41–50.
[5] C. Blaschke, M.A. Andrade, C. Ouzounis, A. Valencia, Automatic extraction of biological information from scientific text: protein–protein interactions, in: Proceedings of the International Conference on Intelligent Systems for Molecular Biology, 1999, pp. 60–67.
[6] D. Appelt, J. Hobbs, J. Bear, D. Israel, M. Kameyama, M. Tyson, FASTUS: a finite state processor for information extraction from real world text, in: Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI), 1993, pp. 1172–1178.
[7] H. Kamp, U. Reyle, From Discourse to Logic, Kluwer, 1993.
[8] R.A. van der Sandt, Presupposition: projection as anaphora resolution, Journal of Semantics (9) (1992) 333–377.
[9] J. Bos, P. Buitelaar, M. Mineur, Bridging as coercive accomodation, in: E. Klein, S. Manandhar, W. Nutt, J. Siekmann (Eds.), Working Notes of the Edinburgh Conference on Computational Logic and Natural Language Processing (CLNLP-95), 1995.
[10] H. Kamp, U. Reyle, A calculus for first order discourse representation structures, Journal of Logic, Language, and Information 5 (1996) 297–348.
[11] H. Schmid, Unsupervised learning of period disambiguation for tokenisation, Tech. rep., Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, 2000.
[12] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence database and its supplement TrEMBL, Nucleic Acid Research (28) (2000) 45–48.
[13] J. Brecher, A practical approach to the sorry state of real-life chemical nomenclature, Journal of Chemical Information and Computer Science 39 (6) (1999) 943–950.
[14] C. Gerstenberger, Semantische Analyse von Namen organischer Verbindungen, University of Stuttgart, 2001.
[15] H. Schmid, Probabilistic part-of-speech tagging using decision trees, in: D. Jones, H. Somers (Eds.), New Methods in Language Processing, Studies in Computational Linguistics, UCL Press, London, GB, 1997, pp. 154–164.

[16] G. Eriksson, K. Franzén, F. Olsson, L. Asker, P. Linden, Exploiting syntax when detecting protein names in text, in: Proceedings of the EFMI Workshop on Natural Language Processing in Biomedical Applications, 1999.
[17] K. Fukuda, T. Tsunoda, A. Tamura, T. Takagi, Towards information extraction: Identifying protein names from biological papers, in: Proceedings of the 3rd Pacific Symposium on Biocomputing, 1998, pp. 707–718.
[18] G. Nenadic, I. Spasic, S. Ananiadou, Terminology-driven mining of biomedical literature, Bioinformatics 19 (8) (2003) 938–943.
[19] S. Abney, Partial parsing via finite-state cascades, in: Proceedings of the ESSLLI'96 Robust Parsing Workshop, 1996.
[20] J. Saric, L. Jensen, R. Ouzounova, I. Rojas, P. Bork, Extraction of regulatory gene expression networks from PubMed, in: Proceedings of the 42nd Meeting of the Association for Computational Linguistics, 2004.
[21] A. Joshi, Y. Schabes, Tree-adjoining grammars, in: Handbook of Formal Languages, vol. 3, Springer, 1997, pp. 69–124.
[22] P. Cimiano, U. Reyle, Ontology-based semantic construction, underspecification and disambiguation, in: Proceedings of the Prospects and Advances in the Syntax-Semantic Interface Workshop, 2003.
[23] D. Davidson, The logical form of action sentences, in: N. Rescher (Ed.), The Logic of Decision and Action, University of Pittsburgh Press, 1967.
[24] U. Reyle, Dealing with ambiguities by underspecification: construction, representation and deduction, Journal of Semantics 10 (2) (1993) 123–179.
[25] R. Muskens, Talking about trees and truth-conditions, Journal of Logic, Language and Information 10 (4) (2001) 417–455.
[26] P. Cimiano, U. Reyle, Talking about trees, scope and concepts, Technical Report, Institute AIFB, University of Karlsruhe, 2004.
[27] P. Dekker, Existencial disclosure, Linguistics and Philosophy (16) (1993) 561–587.
[28] H. Schmid, Lopar: design and implementation, in: Arbeitspapiere des Sonderforschungsbereiches 340, no. 149, 2000.
[29] P. Resnik, Selectional preference and sense disambiguation, in: Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? 1997.
[30] P. Cimiano, ORAKEL: A natural language interface to an *F*-Logic knowledge base, in: Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems, 2004.
[31] N. Asher, A. Lascarides, Bridging, Journal of Semantics 15 (1999) 83–113.
[32] H. Clark, Bridging, in: P. Johnson-Laird, P. Wason (Eds.), Thinking, Readings in Cognitive Science, Cambridge University Press, 1977, pp. 411–420.
[33] T. Gruber, Toward principles for the design of ontologies used for knowledge sharing, in: Formal Analysis in Conceptual Analysis and Knowledge Representation, Kluwer, 1993.
[34] P. Cimiano, On the resolution of bridging references within information extraction systems, ms, University of Stuttgart, 2002.
[35] A. Lascarides, N. Asher, Discourse relations and defeasible knowledge, in: Meeting of the Association for Computational Linguistics, 1991, pp. 55–62.
[36] E. Ratsch, J. Schultz, J. Saric, P. Cimiano, U. Wittig, U. Reyle, I. Rojas, Developing a protein interactions ontology, Comparative and Functional Genomics 4 (1) (2003) 85–89.
[37] P. Blackburn, J. Bos, Representation and Inference for Natural Language, A First Course in Computational Semantics Volume II: Working with Discourse Representation Structures, 1999.
[38] P. Cimiano, Ontology driven resolution of bridging references, in: Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5), 2003.
[39] A. Lascarides, N. Asher, J. Oberlander, Inferring discourse relations in context, in: H.S. Thompson (Ed.), Proceedings of the 30th Annual Meeting of the ACL, Morgan Kaufmann, 1992, pp. 1–8.
[40] C. Müller, M. Strube, Annotating anaphoric and bridging relations with MMAX, in: Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue, 2001, pp. 90–95.
[41] J. Carletta, Assessing agreement on classification tasks: the kappa statistic, Computational Linguistics 22 (2) (1996) 249–254.
[42] J. Landis, G. Koch, The measurement of observer agreement for categorial data, Biometrics 33 (1977) 159–174.

[43] G. Soderland, Building a machine learning based text understanding system, in: Proceedings of the IJCA-2001 Workshop on Adaptive Text Extraction and Mining, 2001.
[44] S. Huttunen, R. Yangarber, R. Grishman, Diversity of scenarios in information extraction, in: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002), 2002, pp. 1443–1450.
[45] The Gene Ontology Consortium, Creating the gene ontology resource: design and implementation, Genome Research (11) (2001) 1425–1433.
[46] A. Rzhetsky, T. Koike, S. Kalachikov, S.M. Gomez, M. Krauthhammer, S.H. Kaplan, P. Kra, J.J. Russo, C. Friedman, A knowledge model for analysis and simulation of regulatory networks, Bioinformatics Ontology 16 (12) (2000) 1120–1128.
[47] P.D. Karp, An ontology for biological function based on molecular interactions, Bioinformatics Ontology 16 (3) (2000) 269–285.
[48] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, University of Sheffield: Description of the LaSIE-II system as used for MUC-7, in: Proceedings of the MUC-7 Workshop, 1998.
[49] R. Yangarber, R. Grishman, Nyu: Description of the Proteus/PET system as used for MUC-7 st, in: Proceedings of the MUC-7 Workshop, 1998.

Philipp Cimiano is a Ph.D. Student at the Institute for Applied Computer Science and Formal Methods (AIFB) at the University of Karlsruhe. He graduated at the University of Stuttgart in Computer Science with a special focus on Computational Linguistics. His research interests include text understanding, computational semantics, information extraction and ontology learning/knowledge acquisition from text.

Uwe Reyle is Professor for Computational Linguistics at the Institute for Natural Language Processing, University of Stuttgart. He studied mathematics and linguistics. His research interests include text understanding, computational semantics of natural language, information extraction as well as ontologies for linguistics and for biochemistry. He is coauthor of the book 'From Discourse to Logic' and coeditor of several books in the area of logic, language and computational linguistics.

Jasmin Šarić studied Linguistics, Computer Science and Psychology at Tuebingen University and graduated in 1999 with a Master's thesis on "Automata theory and Natural Language Quantifiers". He then worked as research associate in the project "Computational Semantics" at the Institute for Natural Language Processing (University of Stuttgart) mainly in the field of temporal and underspecified lexical semantics. At the moment he's writing his Ph.D. thesis on information extraction for biology and ontologies for biology as an associate member of the graduate program of the University of Stuttgart. At EML Research gGmbH he is mainly responsible for the GenIE (Genome Information Extraction) project, which is carried out in cooperation with the University of Stuttgart.