Ontology-Driven Discourse Analysis in GenIE *

Philipp Cimiano
Institute AIFB
University of Karlsruhe
cimiano@aifb.uni-karlsruhe.de

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[1], 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.

1 Introduction

The certainly most important source of biochemical data is the fast growing number of articles available in electronic form. Medline[2] for example contains over 10 million abstracts and approximately 40.000 are added each month. Other important resources are the Journal of Biological Chemistry[3] with more than 50.000 pages published per year as well as the Swiss-Prot database [4], 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 and feed databases with them. 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 ([MNS02]). These target knowledge structures are defined by a given ontology

* The research presented here was conducted at the Institute for Computational Linguistics (IMS) at the University of Stuttgart as well as at the European Media Lab (EML) in Heidelberg.

[1] See http://www.eml.org/english/Research/sdbv/3 for details about the project.

[2] http://www.ncbi.nlm.nih.gov/PubMed/

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

[4] At the time of writing it contained 122.564 protein sequence entries (http://us-expasy.org/sprot/sprot-top.html)

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. In particular, some researches have focused on the extraction of events, i.e. the dynamic aspects of the domain in question ([RRH00], [PCZ02], [YTM01], [RS01], [BAOV99]).
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 protein) 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 becomes feasible to create a conceptual model of the domain which can be exploited within such an approach.
This paper presents a knowledge-based approach to discourse analysis which on the basis of a given ontology and a semantic representation of events computes relations between them that are predefined in the ontology representing a model of the domain in question.
The structure of this paper is as follows: section 2 describes the corpus used and section 3 presents the ontology-driven approach to discourse analysis. Section 4 presents the evaluation of the system against a set of short texts describing the function of proteins taken from the Swiss-Prot database. Finally, section 5 discusses related work and section 6 concludes the paper.

2 The Swiss-Prot-Corpus

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 20189 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, the author 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 author's point of view it denotes rather a state than an event. (See [KR93] 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 of this 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*[5] will be regarded the non-identity 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. *Role*(e_2, e_1), will be called *event role*. Here is an example:

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

The author 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

[5] This nomenclature is introduced by analogy to the famous *bridging* phenomenon ([AL99], [Cla77])

type	occurrences
event coreference	27 (5.1%)
event bridge	137 (25.9%)
event role	28 (5.3%)
total binding events	528 (100%)

Tabelle 1: Results of the classification of binding events as antecedents

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 the author and assigned a unique identifier. Table 2 gives some statistics about the training 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.

	training corpus	test corpus
#tokens	12666	12180
#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

Tabelle 2: Statistics of the training and test corpora

3 The Ontology-Driven Approach

The ontology-driven approach presented in this paper makes use of a semantic representation of the text to make contextual information explicit as well as of a model of the domain in form of an ontology 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 semantic representation language used is DRT ([KR93]). In brief, Discourse Representation Theory (DRT) is a semantic theory in which each sentence gets assigned a Discourse Representation Structure (DRS), i.e. a logical representation of its content. This new DRS is then merged with the DRS representing the discourse processed so far to yield an overall interpretation of a text ([KR93]). In the author's view DRT is so suitable for the task at hand because:

1. DRT has proved so valuable for discourse representation and the analysis of discourse phenomena such as pronoun resolution ([KR93]), presupposition projection ([vdS92]) and bridging ([BBM95]), just to name a very few.
2. In contrast to the traditional template representations used in IE, DRT comes with a well-defined model-theoretic semantics ([KR93]).
3. There already exists a sound and complete calculus for first-order Discourse Representation Structures (DRSs) ([KR96]) which can be used to define an inference mechanism on DRSs.

The approach presented in this paper is inspired in Bos et al's "Bridging as Coercive Accommodation" approach to the resolution of bridging references. In fact, in line with [BBM95], referring or anaphoric expressions are represented by α-marked DRSs[6] 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 ([Cla77]). The most important difference, however, is that it makes use of an ontology of events replacing Bos et al's *qualia structure* ([BBM95]). In contrast to the qualia structure, the ontology does not only represent lexical knowledge, but complex world knowledge about events.
An ontology is a specification of a conceptualization ([Gru93]). 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 ([Gru93]). On the basis of such an ontology, [Cim03] defines a notion of specialization $<_O$ between DRSs. The concepts in C are represented as DRT-based predicate argument structures. A binding between two

[6] These are DRSs marked as unresolved, i.e. which have to be resolved with regard to the preceeding context ([BBM95]).

proteins will for example be represented as follows: $\begin{array}{|l|}\hline e,p_1,p_2 \\ \hline \mathit{bind}(e,p_1,p_2) \\ \mathit{protein}(p_1) \\ \mathit{protein}(p_2) \\ \hline \end{array}$.[7]

A protein-binding event is for example (partially) defined as producing a complex of the involved proteins as a result. Here are the definitions for the resolution of the three different relations considered:

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. $K_1 \leq_O^* K_1'$, and furthermore K_2 is suitable to K_1', where $\leq_O^*$ is the reflexive and transitive closure of $<_O$ and suitability *defines a homomorphism*[8] *on DRSs as in [Cim03].*

The above 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 $[K_1' \oplus R(K_1, K_1')]$ follows logically from K_1 with regard to the ontology as defined in ([Cim03]), i.e. $K_2 \leq_O^* K_2'$ and $K_1 \Longrightarrow_O [K_1' \oplus R(K_1, K_1')]$ and K_2' is suitable to K_1', where $\oplus$ is the merging operator for DRSs ([KR93]).

With the above definition and if the binding of a protein to DNA is defined as leading to the *control/regulation* of the transcription of a certain gene and an *inhibition* is understood as more special than a *control/regulation* (compare figure 1), 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 Role(e_2,e_1) iff K_2 matches a (real) subset of the conditions of K_1, i.e. K_2 is suitable to K_1', where $Con(K_1') \subset Con(K_1)$ and $Con(K)$ are the conditions of the DRS K as defined in ([KR93]).

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

[7] The constants in the upper part of the DRS for the protein-binding event are called *discourse referents* and represent the logical entities appearing in the text. The predicates in the lower part of the DRS represent the so called *conditions* and can be understood as constraints on the possible interpretation of the discourse referents.

[8] Homomorphism is understood here in a mathematical sense, i.e. as a group preserving operation on DRSs.

here, reference resolution is made determinate by choosing the most recent antecedent and minimizing reasoning complexity with regard to the ontology ([Cim03]).

4 Evaluation

4.1 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: $Result$ ([LAO92]) and $Explanation/Elaboration$, where the latter is defined as the disjunction of the Explanation and Elaboration relations considered by Lascarides et al. ([LAO92]). 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.

4.2 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. ([MS01]). The relevant events, states and entities had been previously marked by the author 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 the author, 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 ([Car96]). The overall kappa coefficient has been determined to K=0.31. Following the classification by Landis et Koch ([LK77]) 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 ([Car96]), 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.

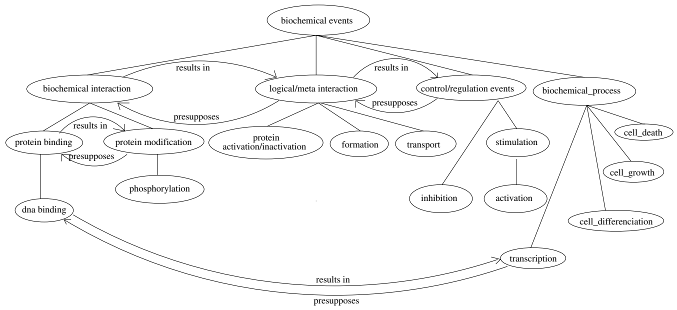

Abbildung 1: Top level of the ontology O_{Bio}

4.3 'Training' the Ontology

The ontology $O_{Bio} = (C_{Bio}, T_{Bio}, D_{Bio})$ has been 'trained' by the author 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 author. These conceptual representations constitute the set C_{Bio}. Furthermore, 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 in form of logical axioms. These axioms form the set D_{Bio} consisting of partial definitions of concepts. The logical language used to represent the axioms as well as taxonomic relations is first-order logic.
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 ($|C_{Bio}|$), 50 taxonomic relations ($|T_{Bio}|$) and 19 axiomatic definitions of concepts ($|D_{Bio}|$). Figure 1 shows graphically the top concepts, the basic taxonomy as well as some of the relations in the ontology.[9] After developing the ontology, the marked events, states and entities of both the training and test corpus have been manually mapped by the author 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, the author 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. ([RSS$^+$03]).

[9]The complete ontology can be found in ([Cim02]).

4.4 Results

The performance of the approach outlined in section 3 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:

$$R = \frac{\text{\#correct answers given}}{\text{\#total correct answers}} \tag{4}$$

$$P = \frac{\text{\#correct answers given}}{\text{\#total answers given}} \tag{5}$$

The *F-measure* is a metric which combines recall and precision into a single value using the formula:

$$F = \frac{(\beta^2 + 1.0) * P * R}{\beta^2 * P + R} \tag{6}$$

where β is the relative weight given to recall over precision. Within the work presented here all F-measures have been calculated using $\beta = 1.0$, i.e. giving equal weight to P and R.

The approach described in section 3 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 and

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%

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

could actually be correct. These observations lead the author 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.

4.5 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 particlar the case for conjunctions such as *by*, *thus*, *because*, *also*, just to name a few. Take for instance the following example:

(7) ALPHA-CONOTOXINS ACT ON POSTSYNAPTIC MEMBRANES, THEY BIND TO THE NICOTINIC ACETYLCHOLINE RECEPTORS (NACHR) AND <u>THUS</u> INHIBIT THEM.

This observation has lead to the idea that discourse relations could also be lexically inferred.[10] 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, i.e.

$$\forall e_1, e_2 \; connect(e_1, e_2, '' thus'') \rightarrow Result(e_1, e_2) \tag{8}$$

In this sense, a lexicon containing conjunctions as well as the corresponding discourse relation which can be 'lexically' inferred from them has been built. Then the test corpus

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

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%

Tabelle 4: Results of the combination of the ontology-driven and the lexically driven approach

has been annotated with the above mentioned *connect*-predicates using the conjunctions specified in this lexicon. Furthermore, a simple approach has been developed which, given a specific instance of a *connect*-predicate specifying the conjunction linking two events together, 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 the following:[11]

$$C = A \cup B - inconsistent(A, B) \tag{9}$$

where $inconsistent(A, B)$ 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 4 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.

[11] Different combination strategies have been explored; the one which worked best is presented here.

5 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. ([AHB+93]). However, recent work in IE ([Sod01], [HYG02], [RS01]) 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. ([HYG02]) 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 ([HYG02]). 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 ([Con01], [RKK+00], [Kar00], [RSS+03]) 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. ([HGA+98]) and Yangarber et al. ([YG98]) 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 ([HGA+98]) 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% ([YG98]).
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.

6 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. Further work will address the syntax-semantic interface, i.e. the mapping from syntactic structures to a DRT-based representation of ontological concepts. 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.

Acknowledgments The author would like to thank Uwe Reyle for fruitful discussions concerning the work presented here as well as Jasmin Saric, Siegfried Handschuh and the anonymous reviewers of the NLDB conference for comments on the first version of this paper. He 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 the author's biochemical questions. Furthermore, he is 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, he would like to acknowledge the European Media Lab and in particular Isabel Rojas for kindly allowing him to use its facilities.

Literatur

[AHB+93] D. Appelt, J. Hobbs, J. Bear, D. Israel, M. Kameyama, and 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.

[AL99] N. Asher and A. Lascarides. Bridging. *Journal of Semantics*, 15:83–113, 1999.

[BAOV99] C. Blaschke, M. A. Andrade, C. Ouzounis, and 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.

[BBM95] J. Bos, P. Buitelaar, and M. Mineur. Bridging as Coercive Accomodation. In E. Klein, S. Manandhar, W. Nutt, and J. Siekmann, editors, *Working Notes of the Edinburgh Conference on Computational Logic and Natural Language Processing (CLNLP-95)*, 1995.

[Car96] J. Carletta. Asessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[Cim02] P. Cimiano. On the Resolution of Bridging References within Information Extraction Systems. Ms. University of Stuttgart, 2002.

[Cim03] P. Cimiano. Ontology Driven Resolution of Bridging References. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5)*, 2003.

[Cla77] H. Clark. Bridging. In P.N. Johnson-Laird and P.C. Wason, editors, *Thinking, Readings in Cognitive Science*, pages 411–420. Cambridge University Press, 1977.

[Con01] The Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. *Genome Research*, (11):1425–1433, 2001.

[Gru93] T.R. Gruber. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. In *Formal Analysis in Conceptual Analysis and Knowledge Representation*. Kluwer, 1993.

[HGA+98] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, and B. Mitchell. University of Sheffield: Description of the LaSIE-II System as used for MUC-7. In *Proceedings of the MUC-7 Workshop*, 1998.

[HYG02] S. Huttunen, R. Yangarber, and R. Grishman. Diversity of Scenarios in Information Extraction. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, 2002.

[Kar00] P. D. Karp. An ontology for biological function based on molecular interactions. *Bioinformatics Ontology*, 16(3):269–285, 2000.

[KR93] H. Kamp and U. Reyle. *From Discourse to Logic*. Kluwer, 1993.

[KR96] H. Kamp and U. Reyle. A Calculus for First Order Discourse Representation Structures. *Journal of Logic, Language, and Information*, 5:297–348, 1996.

[LAO92] A. Lascarides, N. Asher, and J. Oberlander. Inferring Discourse Relations in Context. In H. S. Thompson, editor, *Proceedings of the 30th Annual Meeting of the ACL*, pages 1–8. Morgan Kaufmann, 1992.

[LK77] J.R. Landis and G. Koch. The measurement of observer agreement for categorial data. *Biometrics*, 33:159–174, 1977.

[MNS02] A. Maedche, G. Neumann, and S. Staab. Bootstrapping an Ontology-Based Information Extraction System. In J. Kacprzyk, J. Segovia, P.S. Szczepaniak, and L.A. Zadeh, editors, *Studies in Fuzziness and Soft Computing, Intelligent Exploration of the Web*. Springer, 2002.

[MS01] C. Müller and M. Strube. Annotating anaphoric and bridging relations with MMAX. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, pages 90–95, 2001.

[PCZ02] J. Pustejovsky, J. Castaño, and J. Zhang. Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'02)*, 2002.

[RKK+00] A. Rzhetsky, T. Koike, S. Kalachikov, S. M. Gomez, M. Krauthhammer, S. H. Kaplan, P. Kra, J. J. Russo, and C. Friedman. A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics Ontology*, 16(12):1120–1128, 2000.

[RRH00] T. C. Rindflesch, J. V. Rajan, and L. Hunter. Extracting Molecular Binding Relationships from Biomedical Text. In *Proceedings of the ANLP-NAACL 2000*, pages 188–195, 2000.

[RS01] U. Reyle and J. Saric. Ontology Driven Information Extraction. In *Proceedings of the 19th Twente Workshop on Language Technology*. University of Twente, 2001.

[RSS+03] E. Ratsch, J. Schultz, J. Saric, P. Cimiano, U. Wittig, U. Reyle, and I. Rojas. Developing a protein interactions ontology. *Comparative and Functional Genomics*, 4(1):85–89, 2003.

[Sod01] G.S. Soderland. Building a Machine Learning Based Text Understanding System. In *Proceedings of the IJCA-2001 Workshop on Adaptive Text Extraction and Mining*, 2001.

[vdS92] R. A. van der Sandt. Presupposition: Projection as Anaphora Resolution. *Journal of Semantics*, (9):333–377, 1992.

[YG98] R. Yangarber and R. Grishman. NYU: Description of the Proteus/PET System as Used for MUC-7 ST. In *Proceedings of the MUC-7 Workshop*, 1998.

[YTM01] A. Yakushiji, Y. Tateisi, and Y. Miyao. Event Extraction from Biomedical Papers Using a Full Parser. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'01)*, 2001.