TWLT19
Information Extraction in Molecular Biology

PROCEEDINGS OF THE NINETEENTH
TWENTE WORKSHOP ON LANGUAGE TECHNOLOGY

ESF SCIENTIFIC PROGRAMME ON INTEGRATED
APPROACHES FOR FUNCTIONAL GENOMICS

NOVEMBER 11-14, 2001
ENSCHEDE, THE NETHERLANDS

**Paul van der Vet, Gertjan van Ommen,
Anton Nijholt and Alfonso Valencia (eds.)**

CIP GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Vet van der P.E., Ommen van G.J.B., Nijholt A., Valencia A.

Information Extraction in Molecular Biology
Proceedings Twente Workshop on Language Technology 19
ESF Scientific Programme on Integrated Approaches for Functional Genomics
P.E. van der Vet, G.J.B. van Ommen, A. Nijholt, A. Valencia (eds.)
Enschede, Universiteit Twente, Faculteit Informatica

ISSN 0929–0672

trefwoorden: functional genomics, bio-informatics, information extraction, text mining, computational linguistics

Book orders:
Ms. C. Bijron
University of Twente
Dept. of Computer Science
P.O. Box 217
NL 7500 AE Enschede
tel: +31 53 4893680 – fax: +31 53 4893503
Email: bijron@cs.utwente.nl

Druk- en bindwerk: Reprografie U.T. Service Centrum, Enschede

Preface

TWLT is an acronym of Twente Workshop(s) on Language Technology. Over the years, the topics covered in this series of workshops have evolved from natural-language theory and technology to other aspects of human-computer interaction. The workshops are organised by the Parlevink Research Project, a language theory and technology project of the Centre of Telematics and Information Technology (CTIT) of the University of Twente, Enschede, the Netherlands. For each workshop, proceedings are published containing papers that were presented. For previous volumes in the series, see the final pages of the present volume or the TWLT-webpage[1].

TWLT19, *Information Extraction for Molecular Biology*, returns to the original charter of TWLT by focussing on an important application of natural-language technology, text mining. It continues the multidisciplinary tradition of TWLT: TWLT19 will bring together biologists, bio-informaticians, computational linguists and computer scientists. It is one of a series of workshops organised in the context of the *Scientific Programme on Integrated Approaches for Functional Genomics*[2] funded by the European Science Foundation (ESF).[3] The funding implied that there is no registration fee and that we could invite a relatively large number of speakers. Additional funding by the Executive Board of the University of Twente covered the present proceedings and some social events.

The rapid pace at which genomic information becomes available has directed the attention to the problem of assigning functions to the genes now known. This is a huge challenge that will take concentrated effort by workers in a number of disciplines. The ESF *Scientific Programme on Integrated Approaches for Functional Genomics* supports this work by sponsoring workshops and tutorials and by funding exchanges of researchers. One of the key observations in this field is that the production of data exceeds the researchers' capacity for processing them by far. Therefore, automated means have to be developed to keep up with the information flood. In the present TWLT19 workshop, *Information Extraction for Molecular Biology*, the focus is on the published literature. Authors of journal articles identify relations between genes, regulators and proteins, and each such relation is a piece of the puzzle. We do not know most pathways, but we do know that more systematic perusal of the available literature will no doubt uncover many more pieces of the puzzle. It is therefore in order to bring together workers in relevant disciplines to exchange ideas, views and approaches. The main goal of this workshop as seen by the organisers is to further informal and formal collaborations between different disciplines in order to promote better use of existing literature. Due to divergent research interests, multidisciplinary research is not always easy. The organisers are happy to report, however, that at least among the participants of this workshop there is sufficient convergence.

The present volume of TWLT Proceedings diverges from previous volumes in that the range of papers presented here represents a small portion of the presentations. Here, there apparently is not yet convergence between disciplines, because disciplinary differences are responsible for the state of affairs. Computer scientists gather papers in a volume of proceedings before the event while biologists write their papers normally after the event for publication in an established journal. Even though we should not say so in the preface for a computer science-style proceedings, the biologists' way has something to recommend itself. Journals are less transient than proceedings and they are indexed, proceedings are not. If we would ever want to undertake a major text-mining effort in computer science, we would be stopped by the simple lack of availability of source material. A similar exercise in biology can get a good start, however, because all journal publishers have been keeping full-text electronic archives of their material for a number of years.

Organising a workshop involves many persons beyond the organising committee. We would like to thank the coordinators of the ESF *Scientific Programme on Integrated Approaches for Functional Genomics*, Mike Taussig and Annette Martin of the Babraham Institute, Cambridge UK: they basically *are* the Programme. We would also like to thank (in alphabetical order) Charlotte Bijron, Hendri Hondorp and Alice Vissers-Schotmeijer: without their help this workshop could not have been held.

Anton Nijholt, Gertjan van Ommen, Alfonso Valencia and Paul van der Vet November 2001

[1] http://parlevink.cs.utwente.nl/Conferences/twltseries.html
[2] www.functionalgenomics.org.uk
[3] www.esf.org

Previous TWLT workshops

Previous TWLT workshops were

TWLT1,	*Tomita's Algorithm: Extensions and Applications*. 22 March, 1991.
TWLT2,	*Linguistic Engineering: Tools and Products*. 20 November, 1991.
TWLT3,	*Connectionism and Natural Language Processing*. 12 and 13 May 1992.
TWLT4,	*Pragmatics in Language Technology*. 23 September, 1992.
TWLT5,	*Natural Language Interfaces*. 3 and 4 June, 1993.
TWLT6,	*Natural Language Parsing*. 16 and 17 December, 1993.
TWLT7,	*Computer-Assisted Language Learning*. 16 and 17 June 1994.
TWLT8,	*Speech and Language Engineering*. 1 and 2 December 1994.
TWLT9,	*Corpus-based Approaches to Dialogue Modelling*. 9 June, 1995.
TWLT10,	*Algebraic Methods in Language Processing*. 6-8 December, 1995.
TWLT11,	*Dialogue Management in Natural Language Systems*. 19-21 June, 1996.
TWLT12,	*Automatic Interpretation and Generation of Verbal Humor*. 11-14 September 1996.
TWLT13,	*Formal Semantics and Pragmatics of Dialogue, Twendial'98*. 13-15 May 1998.
TWLT14,	*Language Technology in Multimedia Information Retrieval*. 7-8 December 1998.
TWLT15,	*Interactions in Virtual Environments*. 19-21 May 1999.
TWLT16,	*Algebraic Methods in Language Processing (AMiLP2000)*. 20-22 May 2000.
TWLT17,	*Learning to Behave: Interacting Agents (CEvoLE1)*. 18-20 Oct 2000.
TWLT18,	*Learning to Behave: Internalising Knowledge (CEvoLE2)*. 22-24 Nov 2000.

For the contents of the previous proceedings, please consult the last pages of this volume.

Sponsors

1

University of Twente
The Netherlands

2

[1] http://www.esf.org
[2] http://www.utwente.nl

Contents

Semantic Induction with EMILE, Opportunities in Bioinformatics

Pieter Adriaans
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
pieter.adriaans@ps.net

Abstract

The production of textual information in the biomedical domain is very large. More than 400.000 papers are added to Medline per year. Recently researchers have started to work on the implementation and use of text mining tools to extract knowledge from these huge data sets. For the larger part these tools employ more traditional text mining techniques like frequency counts, n-gram search etc. This paper briefly describes the possibilities of the use of more advanced grammar induction tools for semantic learning. We introduce the EMILE grammar induction tool. Possibilities to use the tool for semantic learning in specific domains, in particular bioinformatics discussed and illustrated with some results. The value of grammar induction tools over standard text mining solutions lies in their more exhaustive structural analysis of the text. The EMILE tool can be downloaded for research purposes at: http://turing.wins.uva.nl/ pietera/Emile/

Keywords: Grammar Induction, semantic learning, bioinformatics term clustering

1 Introduction

Recently a number of researchers have reported on the use of text mining techniques to extract knowledge from biomedical data (Tanabe (1999) , Craven (1999), Blaschke (1999), Swanson (1997)). Over 10 million documents are stored in Medline, and this data set certainly contains implicit hidden knowledge that would have great value for the research community if it could be made explicit. Currently the text mining tools use a variety of shallow techniques like frequency count, n-gram search and similarity measures to compare documents. It is desirable to have techniques that allow a deeper structural understanding of the texts. Grammar induction tools that compress texts in to partial grammars can be of use here. Building on the seminal work of Haris (1966), and Wolff (1982) a number of operational efficient grammar induction tools have been created recently, amongst other EMILE by Adriaans (2000, 1999, 1992a); Vervoort (2000) and ABL by Zaanen (2001, 2000). These approaches to grammar induction all have their root in Harris' linguistic notion substitutability. Since grammar induction use more exhaustive search techniques than standard text mining approaches, they form potentially a powerful addition to the knowledge extraction capabilities in bioinformatics. Garner et al. at the University of Texas South Western have implemented at text mining tool eTBLAST (http://innovation.swmed.edu/) for the analysis of Medline abstracts using document similarity measures. It is a text similarity engine, which accepts a query and then compares it to a collection of other texts. Currently eTBLAST contains about 10% of the Medline abstracts. Initial experiments to enrich the possibilities of eTBLAST with EMILE capabilities are being excuted.

2 Grammar Induction with EMILE

The EMILE algorithm is developed to learn context-free grammars from text, i.e. a collection of strings. Compared with the amount of research on learning regular languages from strings Oliveira

(2000) relatively little research has been on the induction of context-free or even context-sensitive grammars from text. The practical value of tabula rasa induction of a grammar for a natural language from a text is debatable, since it is easier to deduce such a grammar from a tagged corpus and automatic wordclass tagging is a well-developed craft (Halteren (1999)). Furthermore it is clear that natural languages are not strictly context-free but have context-sensitive elements (especially languages like Dutch and German) and sometimes even aspects of free word order that defies a direct explanation in terms of the Chomsky hierarchy all together (Latin, Warlpiri) (Groenink (1997)). The structure of English as the lingua franca of science is relatively close to the context-free ideal. If one applies EMILE to a text that is rich enough it learns those aspects of the language that can be expressed in the context-free format and simply misses others. EMILE works better on English text than on texts in the Dutch or the German language, although it is still perfectly capable of learning partial grammars for these languages. The real value of grammar induction tools surprisingly seems to lie in their capabilities to learn the semantic structure of a domain. I will illustrate this with an example. The following is a context-free grammar of an elementary infinite language that can be used to reason about simple taxonomies:

sentence → *nounphrase verbphrase*	*noun* → dog
sentence → *sentence* and *sentence*	*noun* → animal
sentence → it is not the case that *sentence*	*noun* → tail
nounphrase → a *noun*	*noun* → bird
nounphrase → name	*noun* → mammal
verbphrase → *verb nounphrase*	*name* → Fido
verb → is	*name* → Tweety
verb → has	

In this language we can express simple facts like 'Tweety is a bird and it is not the case that Tweety is a mammal'. The nonterminals, or grammatical types, of this grammar are sentence, nounphrase, verbphrase, noun, name and verb. In a context-free language such types function as substitution classes. By observing substitution restrictions and substitution relations in a sample of a context-free language one can make an educated guess about the types of a language and from the types one may in some cases deduce the structure of the rules. In this language all nouns can be substituted for each other without damaging the syntactic validity of the sentence (substitution salva beneformatione). Any substitution of a verb for a noun destroys the grammaticality of the sentence. The basic concepts in EMILE are contexts, expressions, and context/expression pairs. A context/expression pair is a sentence split into three parts: 'Tweety (has) a tail'. Here 'has' is called the expression, and 'Tweety (.) a tail' is called a context with 'Tweety' as left-hand side and 'a tail' as right-hand side. Not all context/expression pairs make sense from a grammatical point of view e.g.: 'Tweety is a (bird and Fido is a) dog'. Context/expressions pairs are used to investigate substitutability relations. The learning algorithm of EMILE has two phases. In the first phase contexts and expressions are clustered in a matrix in order to find (traces of) types. In the second phase these types are used to create context-free derivation rules. It is a well-known fact that the full class of context-free grammars cannot be learned from text (Gold (1967)). EMILE learns the class of shallow context-free languages with context- and expression separability. In Adriaans (1999) it is argued that these constraints are reasonable for natural languages. EMILE has been tested on a number of texts, varying from the Bible and the Phaistos Disk to Medline abstracts (Adriaans (2000)).

3 Semantic Induction with EMILE

A characteristic sample for a language is a set of sentences from which a grammar induction algorithm can induce the right grammar with high probability. Texts in natural language that are intended to communicate any meaning are almost never characteristic samples for the language, whatever size they have. This is particularly the case of sets of true sentences that are used to characterize a certain knowledge domain. Most sentences that are syntactically valid are simply not true. Yet we need these sentences in our sample to find the right grammatical types. In a

characteristic sample for the grammar in the previous paragraph we would want to have sentences like 'Fido is Tweety', 'Tweety is Fido', 'A dog is a bird', etc. If this language is used to describe a real taxonomy these sentences are ruled out. This observation forms a serious threat to a research program that aims at learning natural language grammars from texts. Still, this drawback has also an advantage. If EMILE is applied to text samples of descriptive true sentences of a specific knowledge domain it learns a partial semantic 'grammar' that can be seen as an intermediate compression level between the syntax and the language. This fortunate behavior of EMILE is due to the principle of compositionality that is considered by some linguists to be universal for human languages. This principle states that the syntax of a language is an algebra, the semantics is an algebra and there is a homomorphism mapping elements of the syntactic algebra onto elements of the semantic algebra (Partree (1997)). A theory that models the complex interplay between partial syntactic and semantic information in language learning is developed in Adriaans (1992a) and Adriaans (1992b). The principle of compositionality will for our taxonomic language enforce a homomorphism between syntax and semantics. This homomorphism maps the syntactic type sentence onto the set semantic types consisting of true_sentence and false_sentence and the syntactic type noun onto a partitioning of that class in to meaningful semantic subclasses i.e. m_noun and b_noun. A fragment of this more complicated 'semantic grammar' (i.e. algebra) is:

true_sentence → *m_nounphrase m_verbphrase*	*p_verb* → is
true_sentence → *b_nounphrase b_verbphrase*	*m_noun* → dog
m_nounphrase → a *m_noun*	*b_noun* → bird
v_nounphrase → a *v_noun*	(. . .)
m_verbphrase → *p_verb* mammal	
b_verbphrase → *p_verb* bird	
(. . .)	

This semantic grammar describes a set of true sentences like 'A dog is a mammal'. If EMILE has a characteristic sample for this semantic algebra (and the algebra is shallow, with context separability etc.) it will learn this semantic grammar, and thus a taxonomy for the underlying knowledge domain. If syntactic learning can be described as investigation of substitutability salva beneformatione then semantic learning is the investigation of substitutability salva veritate. If the sample is not characteristic EMILE will learn a partial semantics that still might be useful. It has to be mentioned that semantic learning is in general much harder than syntactic learning. Syntax has a tendency to be relatively simple because it is a set of conventions that people use to exchange information. It is governed by principles of efficiency. There is no limit to the complexity of meanings that people want to express in language.

4 Application in the Biomedical Domain

In the previous paragraphs we have argued from a theoretical point of view that grammar induction tools have semantic learning capabilities. In this context they might be useful for text mining purposes in the biomedical domain. At this moment it is not clear whether these tools have real advantages to offer above the shallow text mining solutions that are already used. We did a test with EMILE on 91 PubMed abstracts selected with the keywords 'cancer' and 'polymorphism'. In these abstracts about 35.000 expressions and 40.000 contexts were identified. Among the rules learned some seem to have scientific relevance (the numbers between squared brackets indicate grammatical types):

[11] → *LOH was* [105] %
[105] → **identified** in 13 cases (72
[105] → **detected** in 9 of 87 informative cases (10
[105] → **observed** in 5 (55

Clearly such a data set is too small to converge to useful information. A more ambitious project was a grammatical analysis of *Molecular Biology of the Cell,* 3^{rd} edition. We analyzed about 2.5 megabytes of data (text only). In the final analysis 60% of the text was used. Some statistics are:

Number of different sentences read	5461
Number of different words	13396
Number of different contexts	896343
Number of different expressions	782123
Number of different grammatical types	67
Number of dictionary types	17

In this case the rules that were learned seem to have some scientific relevance, although the collected knowledge is rather ad hoc and sparse. The numbers between squared brackets indicate grammatical types. The type [0] is the sentence type. There is information about specific cells:

[0] → Eucaryotic Cells [14]
[14] → Contain Several Distinctive Organelles
[14] → Depend on Mitochondria for Their Oxidative Metabolism
[14] → Contain a Rich Array of Internal Membranes
[14] → Have a Cytoskeleton

There is some chemical structure learned:

[0] → [1].
[1] → B-OH + ATP - [19] 2
[19] → > B-O-P + ADP
[19] → > B-O-P-P + AMP

[0] → The [2]
[2] → [35] Is Asymmetrical
[35] → DNA Replication Fork
[35] → Lipid Bilayer

There are some relevant biological rules:

[0] → [1].
[1] → This [3]
[3] → phenomenon is [37]
[37] → known as gene conversion
[37] → called genomic imprinting

[0] → [64] in the Golgi Apparatus
[64] → Oligosaccharide Chains Are Processed
[64] → Proteoglycans Are Assembled

[0] → Mitochondria and Chloroplasts Contain [66]
[66] → Complete Genetic Systems
[66] → Tissue-specific Proteins

Although it is too early to draw conclusions about the possibility of real knowledge extraction from texts these results are not unpromising. Analysis of a textbook like *Molecular Biology of the Cell* has several drawbacks. A certain amount of redundancy is necessary to get the learning process going. This form of redundancy is rare in a comprehensive textbook. Also the sentences in such a publication are fairly complex. This makes it more difficult to find relevant phrases. Additional experiments with more data are necessary. Also the possibility to use seed grammars to speed up the learning process is investigated. Such a seed grammar could consist of a taxonomy of the domain in a language close to the one identified in paragraph 2. The terms of such a taxonomy could function as kernels for the clustering of semantically related scientific terms.

5 Conclusion

Grammar induction tools applied to large text corpora in de biomedical domain might prove to be a viable addition to existing text mining applications. Initial tests on biomedical data show the possibility of uncovering implicit knowledge contained in those texts. More research on larger corpora is necessary as well as close cooperation with domain experts in the biomedical field.

References

Harris Z.S. (1966). Structural Linguistics, University of Chicago Press, Chicago, IL, USA and London, UK 7th (1966) edition, 1951.

Adriaans P.W., Trautwein M. and Vervoort M. (2000). Towards High Speed Grammar Induction on Large Text Corpora, in V. Hlavác, K.G. Jeffrey and J. Wiedermann (Eds.), SOFSEM 2000, LNCS 1963, pages 173-186, 2000.

Adriaans P.W. (1999). Learning shallow context-free languages under simple distributions. Technical Report PP-1999-13, Institute for Logic, Language and Computation (ILLC), Amsterdam, 1999.

Adriaans P.W. (1992). Language learning from a categorial perspective, PHD thesis, Universiteit van Amsterdam, Amsterdam, the Netherlands, November 1992.

Vervoort M. (2000). Games, Walks and Grammars. PHD thesis, Universiteit van Amsterdam, Amsterdam, the Netherlands, 2000.

Weeber M. Literature-based Discovery in Biomedicine, PHD thesis, Universiteit van Groningen, Groningen, the Netherlands, 2001.

Zaanen M. van and Adriaans P.W. (2001). Alignment-Based Learning versus EMILE: a Comparison. In B. Kröse et al. BNAIC'01, Proceedings of the 13th Dutch-Belgian Artificial Intelligence Conference, pages 315-322, 2001.

Wolff G.J. (1982). Language acquisition, data compression and generalization, Language and Communication, 2(1): pages 57-89, 1982.

Tanabe L. et al. (1999). MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques 27: pages 1210-1217, 1999

Craven M. and Kumlien J. (1999). Constructing biological knowledge bases by extracting information from text sources. ISMB '99, pages 77-86, 1999.

Blaschke C. et al. (1999). Automatic extraction of biological information from scientific text: protein-protein interactions. ISMB '99, pages 60-67,1999.

Swanson D.R. and Smalheiser N.R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. in Artificial Intelligence 91: pages 183-203, 1997.

Zaanen M. van. (2000). ABL: Alignment Based Learning. In Proceedings of the COLING; Saarbrücken, Germany, pages 961-967, 2000.

Oliveira A.L. (2000). Grammatical Inference: Algorithms and Applications, Springer LNCS 1891, 2000.

Halteren H. van. (1999). Syntactic Wordclass Tagging, Kluwer, Dordrecht, 1999.

Groenink A. (1997). Surface without structure. Word order and tractability issues in natural language analysis, PHD thesis Universiteit Utrecht, Utrecht, the Netherlands, 1997.

Gold E.M. (1967). Identification in the limit. Information and Control, 10, pages 447-474, 1967.

Partree B.H. with H.L.W. Hendriks. (1997). Montague Grammar. In Logic and Language, J. van Benthem and A. ter Meulen (Eds.), (1997)., Elsevier Science Publishers, pages 5-93, 1997.

Adriaans P.W. (1992). A domain theory for categorial language learning algorithms. In Proceedings of the Eighth Amsterdam Colloquium, P. Dekker and Martin Stokhof (eds.), pages 1-16.

A Terminology Management Workbench for Molecular Biology*

Sophia Ananiadou
Computer Science
University of Salford, United Kingdom
S.Ananiadou@salford.ac.uk

Hideki Mima
Dept. of Information Science
University of Tokyo, Japan
mima@is.s.u-tokyo.ac.jp

Goran Nenadić
Computer Science
University of Salford, United Kingdom
G.Nenadic@salford.ac.uk

Abstract

In this paper we introduce the design of a web-based integrated terminology management workbench, in which information extraction and intelligent information retrieval/database access are combined using term-oriented natural language tools. Our work is placed within the BioPath research project whose overall aim is to link information extraction to expressed sequence data validation. The aim of the tool is to extract automatically terms, to cluster them, and to provide efficient access to heterogeneous biological and genomic databases and collections of texts, all wrapped into a user friendly workbench enabling users to use a wide range of textual and non textual resources effortlessly. For the evaluation, automatic term recognition and clustering techniques were applied in a domain of nuclear receptors.

Keywords: terminology management, automatic term recognition, term clustering

1 Introduction

The increasing availability of electronically available texts in molecular biology and biomedicine demands the use of appropriate computer tools that can perform information and knowledge retrieval efficiently. The size of knowledge in molecular biology is increasing so rapidly that it is impossible for any domain expert to assimilate the new knowledge without automated means for knowledge acquisition. Vast amounts of knowledge still remain unexplored and this poses a major handicap to a knowledge intensive discipline like molecular biology or biomedicine.

Information retrieval (IR) either via keywords or via URL links have been used intensively to navigate through the WWW in order to locate relevant knowledge sources (KSs). While URLs can be specified in advance by the domain specialists, like links in hypertexts, IR via keywords can locate relevant knowledge sources on the fly. URLs specified in advance are more effective in locating relevant KSs, but they cannot cope with the dynamic and evolving nature of KSs over the WWW. On the other hand, links using keywords, like in a typical IR system, can certainly cope with the dynamic nature of KSs in the WWW by computing links on the fly, but this technique often lacks the effectiveness of the direct links via URLs, as users are often forced to make tedious trials in order to choose the proper sets of keywords to obtain reasonably restricted sets of KSs. This is a well-known problem of WWW querying techniques and the techniques that combine the advantages of these two approaches are needed. Furthermore, since the URLs are often too coarse to locate relevant pieces of information, users have to go through several stages of information seeking process. After identifying the URLs of the KSs that possibly contain relevant information,

*This research is supported by LION BioScience, http://www.lionbioscience.com

they have to locate the relevant pieces of information inside the KSs by using their own navigation functions. This process is often compounded by the fact that users' retrieval requirements can only be met by combining pieces of information in separate databases (or document collections). The user has to navigate through different systems that provide their own navigation methods, and has to integrate the results by herself/himself. An ideal knowledge-mining aid system should provide a seamless transition between the separate stages of information seeking activities.

The ATRACT system, presented in this paper, aims at this seamless navigation and terminology management for the specific domain of molecular biology. It is 'term-centered', as we assume that documents are semantically characterized by sets of technical terms which should be used for knowledge retrieval. Therefore, the very first problem to address is to recognise terms. However, as the amount of new terms introduced in the domain is increasing on daily basis, we need automatic term recognition methods.

The paper is organised as follows: in section 2 we briefly overview ATRACT, and in section 3 we present the design of the system. In the next section we present an analysis and evaluation of our experiments conducted on corpora in the domain of nuclear receptors. The section 5 concludes the paper.

2 ATRACT: an Integrated Terminology Management Workbench

ATRACT (Automatic Term Recognition and Clustering for Terms) is a part of the ongoing BioPath[1] project (Ananiadou et al., 2000). The goal of the project is to develop software components allowing the investigation and evaluation of cell states on the genetic level according to the information available in public data sources, i.e. databases and literature. The main objective of ATRACT is to help users' knowledge mining by intelligently guiding the users through various knowledge resources and by integrating data and text mining, information extraction, information categorization and knowledge management.

As in traditional keyword based document retrieval systems, we assume that documents are characterized by sets of **terms** which can be used for retrieval. However, we differentiate between index terms and technical terms, and in this paper we are referring to *technical terms* i.e. the linguistic realisation of specialised concepts in the domain. In general, technical terms represent the most important concepts of a document and characterize the document semantically. We also consider contextual information between a term and its context words, since this information is important for improvement of term extraction, term disambiguation and ontology building.

A typical way of navigating through the knowledge resources on the WWW via ATRACT is that a user whose interest is expressed by a set of key terms retrieves a set of documents (e.g. from the MEDLINE database Medline (unknown)). Then, by selecting the terms that appear in the document, s/he retrieves fact data from different databases in the WWW. Which databases have to be accessed should be determined automatically by the system. In order to implement the term-centered navigation described above, we have to deal with the following problems:

Term recognition. In specialized fields, there is an increased amount of new terms that represent newly created concepts. Since existing term dictionaries cannot cover the needs of specialists, automatic term extraction tools are needed for efficient term discovery.

Selection of Databases. There is a multitude of databases accessible over the WWW dealing with biological and genomic information. For one type of information there exist several databases with different naming conventions, organisation and scope. Accessing the relevant database for the type of information we are seeking is one of the cricial problems in molecular biology. Once the suitable database(s) is found, there is the difficulty to discover the query items within the database, as well. In addition, naming conventions in many domains (especially in molecular biology) are highly ambiguous even for fundamental concepts (e.g. *'tumor'* can correspond to either a decease, or the mass of tissue; on the other hand, *'TsaB'* is a protein, and *'tsaB'* is a gene), which effects selection of appropriate databases.

[1] BioPath is a collaborative EUREKA research project coordinated by LION BioScience and ValiGen.

ATRACT aims to provide solutions to the problems described above by integrating the following components: automatic term recognition, context-based automatic term clustering, similarity-based document retrieval, intelligent database access and terminology storage and management.

3 ATRACT System Design

The ATRACT system contains the following components (see figure 1):

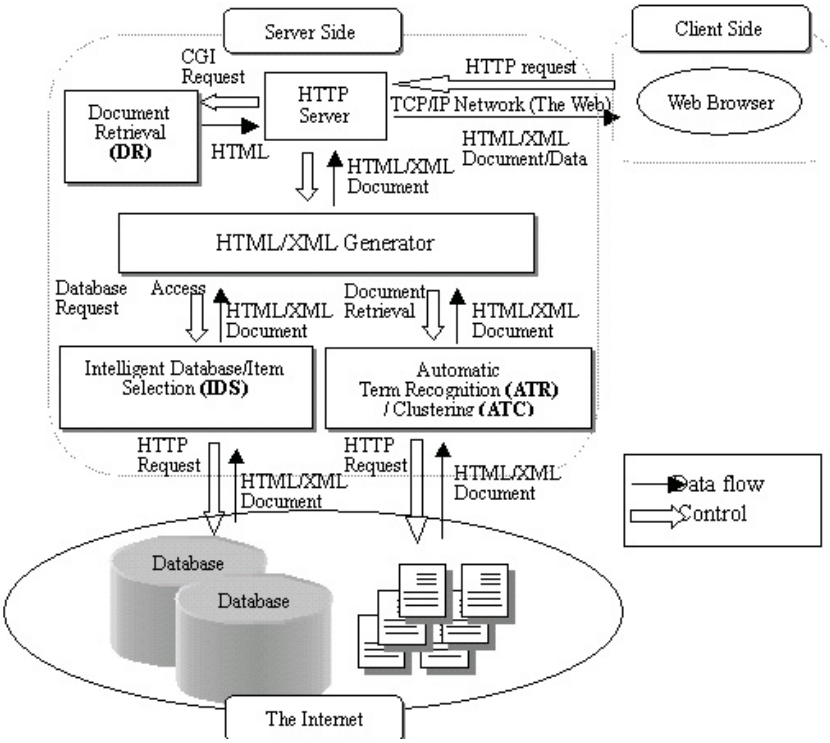


Figure 1: system design of ATRACT

(1) Automatic Term Recognition (ATR). The ATR module recognizes terms included in HTML/XML documents using the *C/NC-value* method Frantzi et al. (2000). The C/NC-value method recognizes term candidates on the fly from texts which often contain unknown or new terms. This method is a hybrid approach, combining linguistic knowledge (term formation patterns) and statistics (frequency of occurrence, string length, etc). C/NC-value extracts multi-word terms and performs particularly well in recognizing nested terms i.e. sub-strings of longer terms. One of the innovative aspects of NC-value (in addition to the core C-value method) is that it is context sensitive. In a specific domain, lexical preferences of the type of context words occurring with terms are observed Maynard et al. (2000), Maynard et al. (2001). The incorporation of contextual information is based on the assumption that lexical selection is constrained in sublanguages and that it is syntactified. The user can experiment with the results of the ATR module by tuning parameters such as threshold value, threshold rate, weights, selection of part-of-speech categories, choice of linguistic filters, number of words included in the context etc. according to his/her specific interests.

(2) Automatic Term Clustering (ATC). Contextual clustering is beneficial for resolving the terminological opacity and polysemy, common in the field of molecular biology. Table 1, for example, shows problems of term ambiguity in the field. The same terms which are fairly specific and domain dependent still have several different meanings, depending on the actual context in

which these terms appear. This means that, depending on the context, we have to refer to different databases to retrieve fact data of these terms.

term	protein	enzyme	compound
amino acid	+	–	–
amino acid sequence	+	–	+
pyruvate dehydrogenase	+	+	+
pyruvate carboxylase	+	+	+

Table 1: term ambiguity

The ATC module classifies terms recognized by the ATR module based on contextual clustering and statistical techniques. It is an indispensable component in our knowledge-mining system, since it is useful for term disambiguation, knowledge acquisition and construction of domain ontology. The approach is based on the observation that terms tend to appear in close proximity with terms belonging to the same semantic family Maynard et al. (2000). If a context word has some contribution towards the determination of a term, there should be a significant correspondence between the meaning of that context word and the meaning of the term. Based on that observation, we compare the semantic similarities of contexts and terms. The clustering technique is based on automatically deriving a thesaurus based on the AMI (Average Mutual Information) hierarchical clustering method Ushioda (1996). This method is a bottom-up clustering technique and is built on the C/NC-value measures. As input, we use bigrams of terms and their context words, and the output is a dendrogram of hierarchical term clusters.

(3) Similarity-based Document Retrieval (DR). DR is a VSM (vector space model)-type document retrieval module. It retrieves texts associated with the current document, allowing the user to retrieve other related documents by assigning selected keywords and/or documents using similarity-based document retrieval. The user can also retrieve documents by specifying keywords.

(4) Intelligent Database/item Selection (IDS) using database (meta-) ontology. IDS selects the most relevant databases and their items using term class information assigned by ATC module and database's (meta-)ontology information. All terms are 'clickable' and dynamically 'linked' to the relevant databases over the Internet. The relevant databases should be dynamically selected according to the terms and the term hierarchy information. The module is implemented as an HTTP server, designed to choose the appropriate database(s) and to focus on the preferred items in the database(s) according to the user's requirements. The most relevant databases are determined automatically by calculating association scores between the term classes and the description of databases (such as meta-data). Furthermore, the retrieved content can be modified in order to focus on the most pertinent items for the user. It is also possible to show similar entries by calculating similarities using the term classes and the domain specific ontology when an exact matched entry is not found.

4 Experiments and Evaluation

We conducted experiments to confirm the feasibility of our proposed workbench. The evaluation was performed on 2,000 MEDLINE abstracts Medline (unknown) in the domain of nuclear receptors. We focused on the quality of automatic term recognition and similarity measure calculation with the use of automatically clustered terms, as all other techniques are based on term extraction.

— **Term recognition.** We have examined the performance of the NC-value method with respect to the overall performance from the viewpoint of precision and recall by 11-point[2] score, while applying it to the same corpus and the correction set to the C-value. The top of the list produced by C-value (the first 20% of extracted candidate terms) was used for the extraction of term context

[2]11-point score indicates that, for example, precision at recall 0.10 is taken to be maximum of precisions at all recall points greater then 0.10.

words, since these show high precision on real terms. We used 30 context words for all the extracted terms in the evaluation, the number been determined empirically.

Figure 2 (left) shows the 11-point precision-recall score of NC-value method in comparison with the corresponding C-value. It can be observed that NC-value increases the precision compared to that of C-value on all the correspond points for recall. Similarly, NC-value increases the precision of term recognition compared to pure frequency of occurrence. Although there is a small drop in precision compared to C-value in some intervals (figure 2, right), NC-value generally increases the concentration of real terms at the top of the list.

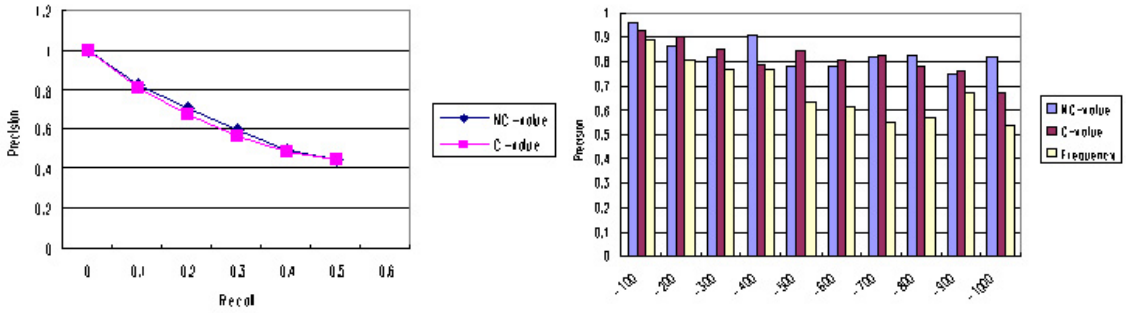


Figure 2: 11-point score (left) and interval precision (right)

— **Clustering terms and database handling.** We used the similarity measure calculation as the central computing mechanism for choosing the most relevant database(s), determining the most preferred item(s) in the database(s), and disambiguating term polysemy. The clustered terms were developed by using Ushioda's AMI-based hierarchical clustering program Ushioda (1996). As training data, we have used 2,000 MEDLINE abstracts. Similarities between terms were calculated according to the hierarchy of the clustered terms. In this experiment, we have adopted a semantic similarity calculation method for measuring the similarity between terms described in Oi et al. (1997). We have used three sets (*DNA*, *PROTEIN*, *SOURCE*) of manually classified terms and calculated the average similarities (*AS*) of every possible combination of the term sets, that is, $AS(X,Y) = \frac{1}{n}\Sigma\ sim(x,y)$, where X and Y indicate each set of the classified terms; $sim(x,y)$ indicates similarity between terms x and y, and n indicates the number of possible combinations of terms in X and Y (except the case where $x = y$). As the table 2 shows, each AS between the same class terms, i.e. $AS(X,X)$, is greater than the others respectively. We believe that it is feasible enough to use automatically clustered terms as the main source of knowledge for calculating similarities between terms.

	DNA	*PROTEIN*	*SOURCE*	# of terms
DNA	**0.533**	–	–	193
PROTEIN	0.254	**0.487**	–	235
SOURCE	0.265	0.251	**0.308**	575

Table 2: average similarities

However, despite these results on clustering and disambiguation, searching for suitable databases on the Web still remains a difficult task. One of the main problems encountered is that we are not certain which databases have items which best describe our request, i.e. we do not know whether the related databases are pertinent to our request. In addition, the required information is sometimes distributed through several databases, i.e. almost all databases are disparate in terms of information contained.

5 Conclusion and Further Research

In this paper, we have presented ATRACT, a web-based integrated terminology management workbench. ATRACT extracts automatically terms based on a combination of linguistic and statistical knowledge, clusters terms and provides seamless navigation and access to heterogeneous databases and collections of texts. The workbench provides a user with a friendly environment for term extraction and clustering from a variety of knowledge and textual sources. The system enables a logical integration of databases on the Web: the design allows the users to refer to the required items gathered from several web databases as if they access certain sophisticated single database virtually.

Important areas of future research will involve improvement of term recognition using semantic/clustered term information and additional syntactical structures (e.g. term variants and coordination Kehagia et al. (2001), Nenadic (2000)), and improvement of database handling. Since our goal is dealing with the 'open world' of databases, due to insufficiency of information on what sort of data is contained in each database, selecting the most associative databases is one of the crucial problems. Therefore we will have to resolve problems of choosing database(s) from large amounts of databases on the Web (and to recognise newly launched databases as well), and to modify the 'view' of each database according to the requirements (since the format styles vary from site to site). We expect that meta-data information could be useful to select database(s) if enough meta-data about each database is available on the Web. Regarding the format style of each database, we expect that the popularity of XML might be a solution of the problem.

References

Ananiadou, S., Albert, S., Schuhmann, D. (2000): "Evaluation of Automatic Term Recognition of Nuclear Receptors from Medline", in Genome Informatics Series, vol.11

Frantzi, K. T., Ananiadou, S., Mima, H. (2000):"Automatic Recognition of Multi-Word Terms: the C-value/NC-value method", in International Journal on Digital Libraries Vol. 3, No. 2, pp.115–130, 2000

Kehagia, K., Ananiadou S. (2001): "Term Variation as an Integrated Part of Automatic Term Extraction", in Proc. of 22nd Conference for Greek Language, Thessaloniki, Greece

Maynard, D., Ananiadou, S. (2000): "Identifying Terms by Their Family and Friends", in Proc. of 18th International Conference on Computational Linguistics, COLING 2000, pp.530–536, Luxembourg

Maynard, D., Ananiadou, S. (2001): "TRUCKS: a Model for Automatic Term Recognition", in Journal of Natural Language Processing, Vol. 8, No. 1, pp.101–125

MEDLINE, National Library of Medicine, `http://www.ncbi.nlm.nih.gov/PubMed/`

Mima, H., Ananiadou, S., Tsujii, J. (1999): "A Web-based Integrated Knowledge Mining Aid System Using Term-oriented Natural Language Processing", in Proc. of The 5th Natural Language Processing Pacific Rim Symposium, NLPRS'99, 13-18

Nenadić, G. (2000): "Local Grammars and Parsing Coordinaton of Nouns in Serbo-Croatian", in Text, Speech and Dialogue - TSD 2000, Lecture Notes in Artificial Intelligence 1902, Springer Verlag

Oi K., Sumita E., Iida H. (1997): "Document Retrieval Method Using Semantic Similarity and Word Sense Disambiguation in Japanese", in Journal of Natural Language Processing, Vol.4, No.3, pp.51-70

Ushioda A. (1996): "Hierarchical Clustering of Words", In Proc. of COLING '96, Copenhagen, Denmark

Biological Function and DNA Expression Arrays

Christian Blaschke
Protein Design Group at the CNB/CSIC
Cantoblanco, Universidad Autonoma, 28049 Madrid, Spain
blaschke@cnb.uam.es

Luis Cornide
ALMA Bioinformatica
28760 Tres Cantos, Spain
lcornide@almabioinfo.com

Juan Carlos Oliveros
Protein Design Group at the CNB/CSIC
Cantoblanco, Universidad Autonoma, 28049 Madrid, Spain
oliveros@cnb.uam.es

Alfonso Valencia
Protein Design Group at the CNB/CSIC
Cantoblanco, Universidad Autonoma, 28049 Madrid, Spain
valencia@cnb.uam.es

Abstract

DNA arrays are one of the types of large-scale experiments that have been developed over the last years. These experiments allow new biological insights but also provide an overwhelming flow of data that has to digested and analyzed properly. We developed an information extraction system (GEISHA) that provides an overview of the literature related to the genes that are implicated in an experiment. It extracts keywords and the most important parts of the related abstracts and re-organizes the information in a way that with much less effort a deeper insight in what was published already is possible. Here we present an overview of the system and the results that were obtained in different studies.

Keywords: Information Extraction, DNA arrays, data analysis, clustering, term frequencies

1 Introduction

In the past few decades, biologists have generated a large amount of data that have been published mainly in biological journals. It is now important to be able to recover as much as possible of this information as it constitutes a precious source of additional information for helping to understand the new genomics and proteomics data. More than 11 million abstracts of such papers are contained in the Medline collection and are available at the NCBI (Medline 2001), from publicly available Medline servers (Dr. Felix 2000), or from commercial distributions (SilverPlatter 2000). This collection will expand considerably once the full text of the publications become accessible over the Web in a generalized way (PubMedCentral 2001, E-bioscience 2001).

1.1 GEISHA

Expression arrays have introduced a paradigmatic change in biology by shifting experimental approaches from single gene studies to genome-level analysis. The first wave of experiments is already available for *Escherichia coli* (Richmond 1999), *Saccharomyces cerevisiae* (Cho 1998; Chu

1998; DeRisi 1997; Eisen 1998; Holstege 1998; Spellman 1998; Wodicka 1997), human (Alizadeh 2000; Iyer 1999) and rat tissues (Wen 1998). Some of these results have been made publicly available (Jennings 1999), stimulating the development of new approaches required for this complex analysis (see Bassett 1999).

The main result of expression array experiments is the discovery of sets of genes with similar gene expression patterns (expression-based gene clusters). The underlying assumption is that these gene clusters are related by their participation in common biological processes (Lockhart 2000). The operations carried out to define the biological meaning of these clusters typically involve consulting functional annotations in different sequence databases such as SWISS-PROT (Bairoch 1997; SwissProt 2001) or other specialized databases, such as YPD (Hodges 1999; Proteome 2001). This information is often insufficient and bibliographic information must be consulted, usually by following the links to selected MEDLINE abstracts provided in some sequence databases. Since only a small fraction of these pointers provide direct information about gene function further references are usually collected by querying Pubmed directly (Medline 2001) with gene names. In practice, analysis of a full experiment can imply thousands of references, making the systematic analysis of the differences between gene groups impractical. This situation will become increasingly complex for experiments referring to larger systems, such as the human genome.

Most of the efforts related with the analysis of DNA array experiments concentrated on the definition of standards for the normalization of the raw data (Quackenbush 2001), the exchange format of this data (GEML 2001), microarray image analysis (ImaGene 1999), primary data management (Ermolaeva 1998; Liao 2000) and cluster analysis (Eisen 1998). Development of methods to extract information about the common biological characteristics of gene clusters has received considerably less attention. There is an obvious need for protocols to summarize vast amounts of data in a comprehensive way, algorithms to select information that could be of use to human experts, and tools to guide them through the analysis. As pointed out by Bassett *et al.* (Bassett 1999) "the ultimate goal is to convert data into information and the information into knowledge".

GEISHA (Gene Expression Information System for Human Analysis, (Blaschke 2001; Oliveros 2000)) is conceptually similar to other statistical approaches, such as that previously developed by Andrade and Valencia Andrade 1998 for the assignment of functional keywords to protein families. The GEISHA system involves the annotation of function for groups of genes that show similar expression patterns in DNA array experiments. First the system uses the groups of genes as a framework for clustering the related literature. In a second step it estimates the frequency of relevant words in the various literature clusters, and then in a third step these frequencies are compared in order to assess their statistical relevance (in the form of Z-scores). A similar procedure is applied to the extraction of complete sentences specific to the various gene clusters.

Since biological information is often expressed in composite terms such as *DNA polymerase* and *RNA polymerase*, these constructions are detected by analyzing the frequency of these co-occurrences in comparison to the expected frequency of the individual component words.

The results of the GEISHA system have been extensively compared to the annotations provided by databases and human experts, showing how in many cases GEISHA was able to extract relevant or alternative information to that provided by other sources.

In addition to GEISHA other approaches have been published that use of the literature in relation with the analysis of DNA arrays.

MedMiner (Tanabe 1999) uses a pre-defined list of keywords which where compiled for different domains in molecular biology and medicine to filter the abstracts returned from a MEDLINE search and to select the sentences that best describe the document. In addition the information of GeneCards (Rebhan 1997) is used to obtain synonyms for the genes specified by the user and to extend the query. This information is presented in web pages which allows a quick overview of the results. It proved to be useful to some extent for the analysis of DNA arrays because the overwhelming amount of text related with the genes in an experiment are easier to handle.

Jenssen *et al.* (Jenssen 2001) constructed a network of gene relations for Human simply by counting co-occurrences of gene symbols obtained from a public repository in MEDLINE abstracts. These relations where then compared to the results obtained by clustering the data from DNA

expression arrays. This simple approach gives very interesting results because genes that are functionally related can have totally different expression patterns (and belong to different clusters). But the cases in which they appear in the same abstracts their relation is not evident in the experiments. This information can be used to propose new experiments.

Shatkay *et al.* (Shatkay 2000) developed a method that detects similar documents to a given seed document. It is not based on a static similarity measure of the word frequencies in the different abstracts but tries to detect the similarity of the "theme" between text and associate abstracts and the query document. As a "side product" keywords for each theme are extracted that serve for the interpretation by the users. The objective is somehow similar to that of Jenssen *et al.* (Jenssen 2001) because genes with the same themes in different clusters point to a relation between groups which where not detected in the experiments. The problem of the methods is that a "kernel document" for each gene has to be selected and no automatic procedure for this was presented by the authors which will limit the application of this method to large scale experiments since the selection of this initial document will influence the results considerably.

2 Methods

GEISHA provides organized functional information about expression array experiments by connecting the information stored in large collections of Medline abstracts with the corresponding gene expression clusters.

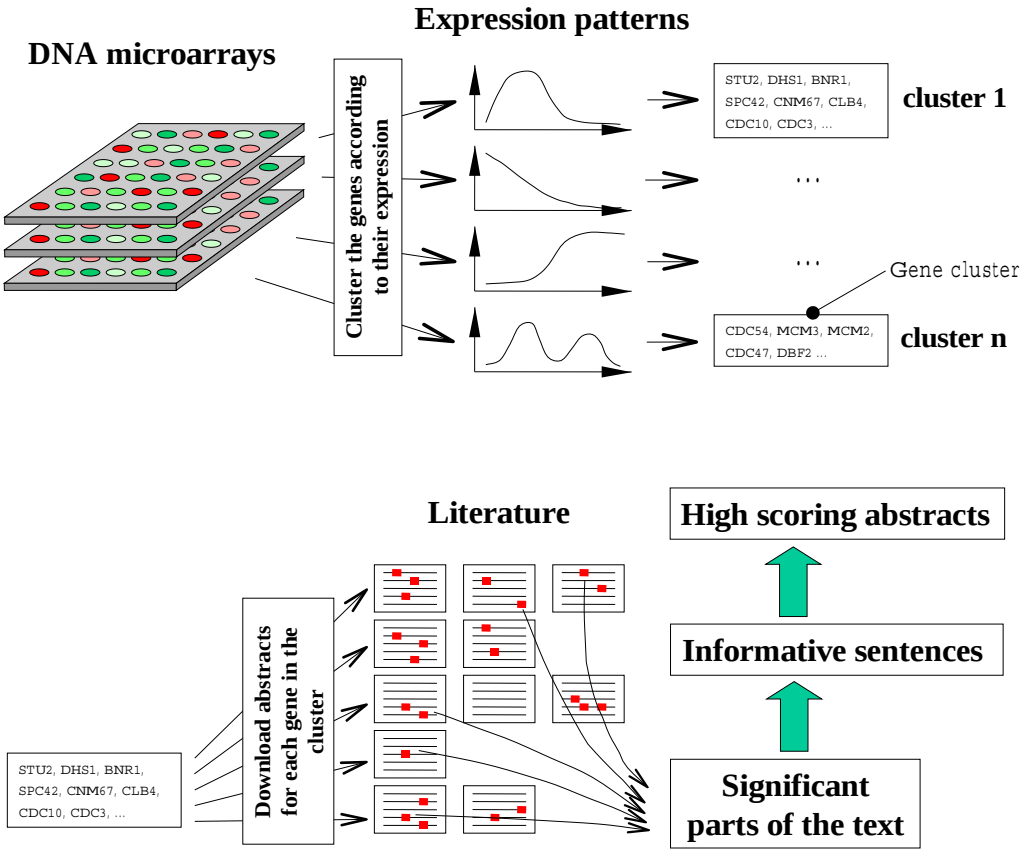


Figure 1: Overview of the GEISHA system.

Figure 1 shows the basic steps in GEISHA assisted DNA expression array analysis. The

genes are grouped according to the similarity in their expression patterns. Then the literature corresponding to each gene is collected for the different clusters and the significant parts of the text are extracted by comparison of the term frequencies in all the clusters. These are used in the consecutive steps to detect highly informative abstracts and to score the abstracts by their information content.

2.1 Implementation and access to the system

ALMATextMiner (the commercial version of GEISHA) is a system that, using information extraction techniques similar to those of GEISHA, helps researchers interpret the results of a DNA array experiment by analyzing the available literature so as to characterize the clusters of genes produced by the experiment. This system has been implemented as a web application and provides a simple and intuitive interface for using the tool and interpreting the results obtained.

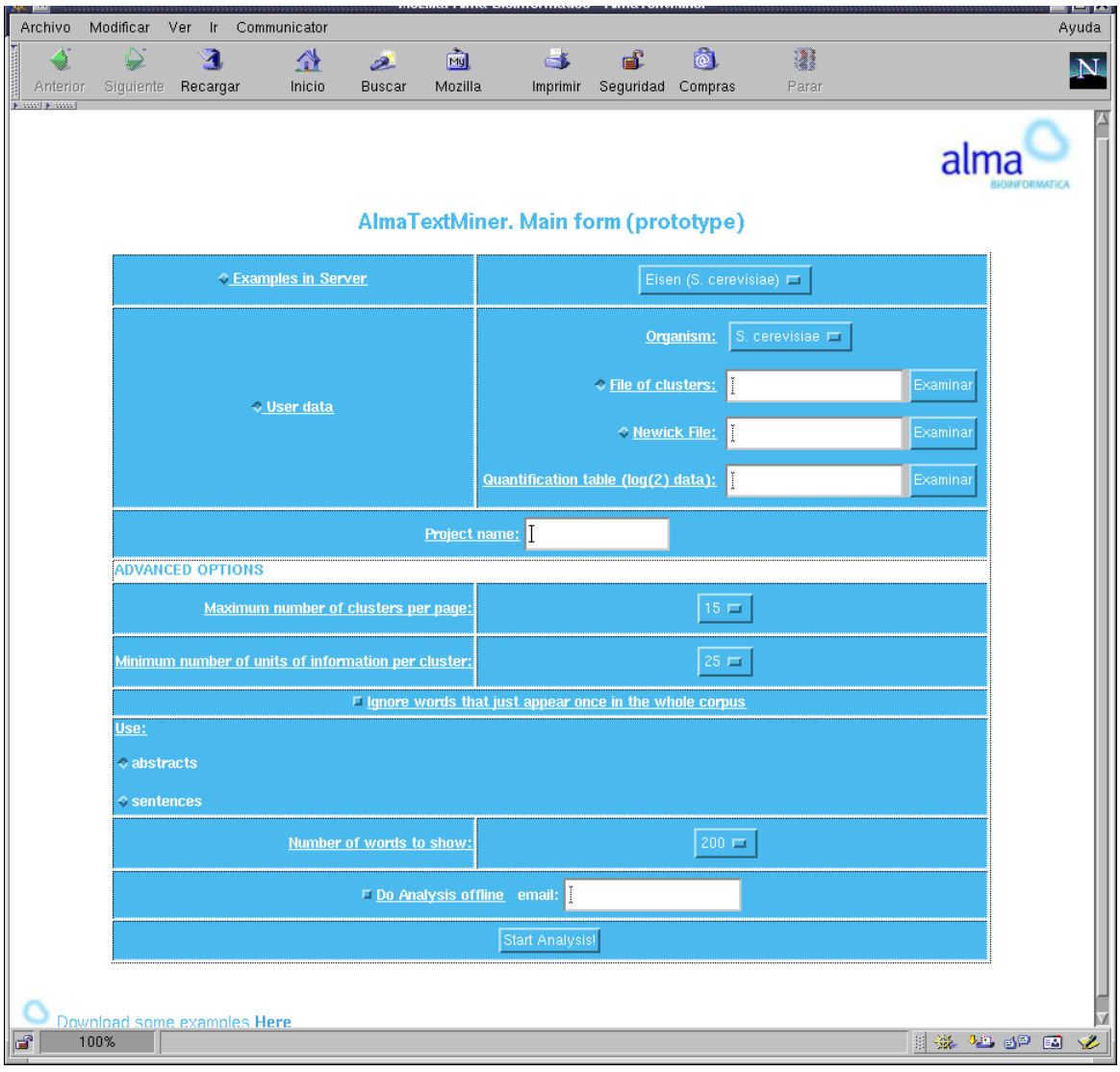


Figure 2: The input page of the ALMATextMiner where the results DNA array experiments are uploaded and the options for the analysis are specified.

The analysis requires an input file to be provided that specifies the composition of the clusters.

A file of the associated expression profiles may also be supplied. As can be seen in Figure 2, a large set of options are then provided for determining how the analysis will be carried out (eg. type of units of information to be used) and how the results will be presented (eg. number of clusters to be shown per HTML page).

ALMATextMiner is currently set up for analysis involving genes from *Saccharomyces cervisiae*, *Escherichia coli* and *Arabidopsis thaliana*, as the system maintains a large database of documentation on these organisms. Analysis of the results of DNA array experiments involving a very large number of clusters and genes can be time-consuming, hence an "offline analysis" option is also included. With this option, the user leaves the server to run the analysis independently and is then notified via e-mail when the analysis terminates.

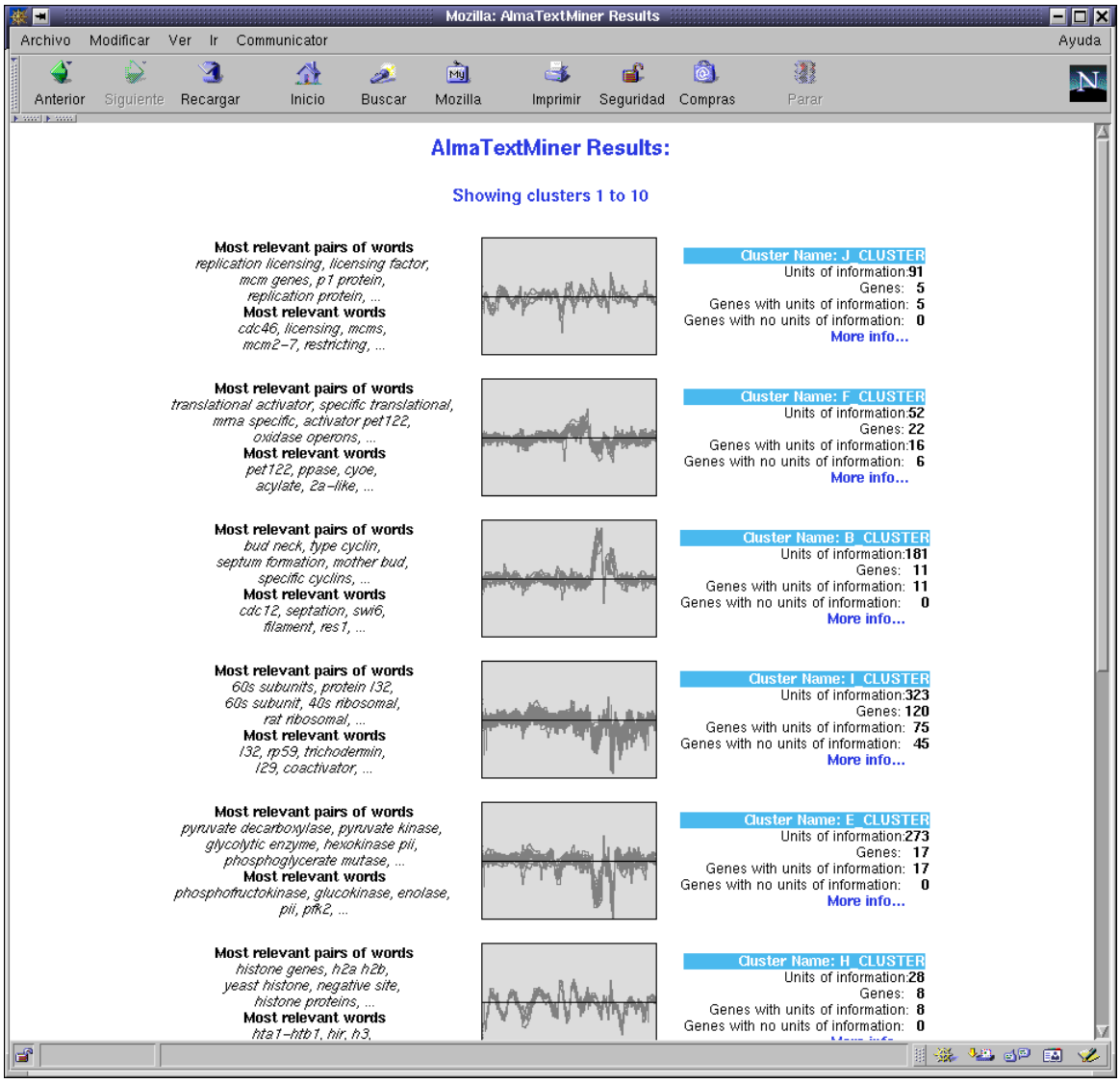


Figure 3: One of the output pages of the ALMATextMiner that shows the expression profiles for the different cluster along with information on the cluster and part of the extracted information from the literature.

Once the analysis process is complete, the results are presented to the user as a series of web pages (Figure 3), where summarized information for each cluster is displayed (e.g. number of genes

in the cluster; number of units of information assigned to the cluster), together with the associated expression profiles (if the file containing these has been supplied).

From these summary pages the user may then access further pages where the full set of information generated for a cluster is displayed, in particular lists of the sentences and single and compound terms that are most characteristic of the cluster. A list of those authors that are most relevant to the cluster is also provided. Lastly, if the file of expression profiles has been provided then a table of these is displayed for those genes making up the cluster, together with a description of the experimental conditions.

These results are stored on the server for several days so that they may be consulted by the user whenever required. More information about the contribution of ALMA to the field of information extraction can be obtained at http://www.almabioinfo.com/techno_infoex_science.html.

2.2 Text corpus

The methodology discussed here was first applied to the yeast expression data published by Eisen *et al.* (1998). These experiments monitored the expression of yeast cells in 79 different experiments including diauxic shift, mitotic cell cycle, sporulation, temperature and reducing shocks. The GEISHA system was applied to the 254 genes that showed important differences in gene expression, corresponding to ten clusters (genes and clusters from Figure 2 in Eisen 1998). As a first step these 10 clusters were analyzed. Based on the encouraging results obtained the data of the original experiments were clustered with a different algorithm based on growing self-organizing maps (Herrero 2001) and analyzed with GEISHA.

At the time of collecting the text corpus, 20,897 Medline abstracts were found that mentioned at least one yeast gene (taking into account synonymous names and gene name + p for the proteins expressed, e.g. cdc47p).

2.2.1 Relating abstracts to gene clusters

The gene clusters (as obtained by Eisen *et al.* (1998) analyzing the experimental expression data) were used by GEISHA to classify entries of the text corpus. Abstracts were linked to a given cluster if they contained the name of any of the genes in the cluster. Some abstracts can be related to more than one cluster if they contain gene names from different groups. This introduces some additional information at the expense of including undesired noise.

2.3 The procedure

The GEISHA process includes the following steps: (1) calculation of the frequency of the terms associated to the different gene groups comparing the Medline abstracts associated to each group of genes, (2) assessment of the significance (Z-score) of the terms associated to each cluster, (3) analysis of the information provided by the co-occurrence of terms, (4) evaluation of the significance of sentences, (5) selection of abstracts based on the score of their terms, and (6) presentation of the results.

2.4 Frequency of terms

The frequency of the terms in the Medline abstracts associated to each cluster is compared to the frequency of these terms in the other clusters:

$$\overline{f^a} = \frac{\sum_{i=1}^{n} f_i^a}{n} \tag{1}$$

$\overline{f^a}$ is the mean frequency of term a over all clusters, f_i^a is the frequency of term a in cluster i and n is the number of clusters. In other words, we quantify the frequency of documents referring to a term and not the number of times that a term appears in a set of abstracts.

A term is considered significant if it appears more frequently in the abstracts associated to the cluster than in abstracts associated to other clusters.

2.5 Significance of terms

The significance is calculated in terms of Z-score, defined as:

$$Z_i^a = \frac{f_i^a - \overline{f^a}}{\sigma_a} \tag{2}$$

where σ_a is the standard deviation of the distribution of the term a:

$$\sigma_a = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(f_i^a - \overline{f^a})^2} \tag{3}$$

where n is the number of clusters. In our analysis a term is taken as significant if its Z-score is 2.00 or more.

Two reasons support the use of Z-scores in this case even if the distributions are not always Gaussian. First, SD can still be considered a good estimator of diversity for non-normal distributions (Mann 1995) and second, the results that we obtain, even in extreme cases, are reasonable. This happens for example when a term does not occur in most clusters (no relation with that function for most of the genes) and only a few clusters contain a large number of abstracts presenting the term. The Z-score for a cluster containing the term will be correctly assigned to a high value, since the frequency of the corresponding term will be considerably high in comparison with the low average value of the distribution, even when it is normalized by the high SD value of the distribution.

2.6 Information context provided by selected sentences

Significant sentences were selected by dividing the sum of scores of the significant terms by the total number of words in the sentence (significant or not). This is an *ad hoc* procedure that in our experience works better than using the number of significant terms with regard to repetitive occurrences (data not shown). This procedure favors short sentences that accumulate significant terms and concrete information. Very short sentences (less than six words) and very large ones (more than 30 words) were explicitly excluded.

2.7 Information context provided by selected abstracts

A similar procedure was implemented for the selection of abstracts containing relevant information. Abstract score is simply calculated by adding the scores for their sentences. This process favors large abstracts containing many significant sentences. The score enables sorting of abstracts by relevance; these best-scoring abstracts are potentially valuable as first candidates for human inspection in the course of analysis of expression array results.

2.8 Outcome of the GEISHA analysis

GEISHA provides information about selected terms, co-occurrence of terms, significant sentences and abstracts in the form of web-pages that allow navigation between the extracted terms, sentences and selected abstracts on the one hand and the functional information provided by the sequence databases (YPD and SwissProt in this case) on the other hand. The most convenient way to use this information is first to check the terms to obtain a general idea about the functions associated to the different gene clusters, then use the database information for detailed description of the function of some of the known genes. Subsequently it will be necessary to look more closely at in sentences and abstracts in those cases in which the database information is considered insufficient. Access to the abstracts is facilitated by the GEISHA scoring scheme. GEISHA also facilitates information for redefining the selection of the text corpus, which can be improved by using the main terms as keywords for the selection of new Medline entries.

The results presented here are a summary of 2 different studies we performed at the Protein Design Group (discussed in greater detail in Blaschke 2001 and Oliveros 2000) and are accessible at http://montblanc.cnb.uam.es/geisha/ and http://montblanc.cnb.uam.es/SOTAandGEISHA/.

3 Results

3.1 Analysis of key terms for gene clusters

Table 1: J cluster terms and their classification for analysis.

Functional groups	Terms
Minichromosome maintenance	mcm3, mcm2, mcm, mcm4, mcm2 mcm3, mcm5 cdc46, mcm proteins, mcm family, mcm genes, minichromosome, maintenance, minichromosome maintenance, maintenance mcm, mis5, chromosome loss
DNA synthesis	Licensing factor, replicate, replication, replication licensing, replication origins, autonomously replicating, DNA replication, DNA synthesis, S-phase, S phase
Phosphorylation	Protein kinase, dbf2, phosphorylate
Cell cycle	cdc46, cdc47, cdc21, cdc54, cell cycle
Non-specific (biological)	Genetically, nucleus, nuclei, homologues, DNA, phase, m, eukaryote, antibody, mouse, fission, cycle, temperatures, per cell, budding yeast, protein family, protein complex, fission yeast, Schizosaccharomyces pombe, egg extracts, Xenopus egg, hela cells
Non-biological	Once, origin, initiation, throughout, of, early, per, physically, family, member, degree, loss, after, late, play, apparently, implicate, share, associated, localization, non-permissive, progression, detect, raised against, accompanied by, depends on, degrees C, rather than, license

All the terms with a significant Z-score are displayed and grouped by hand (for more details see http://montblanc.cnb.uam.es/geisha/).

The results obtained for one gene cluster (cluster J in Eisen 1998) illustrate the quality of the terms extracted by GEISHA (Table 1). This cluster includes genes related to DNA replication initiation and entrance into cell cycle, including cell division control (CDC) genes such as cdc47 and cdc54, genes related to minichromosome maintenance (mcm2 and mcm3) and dbf2, a protein kinase related to cell division. The terms extracted by GEISHA can be classified by manual inspection into four different functions: minichromosome maintenance, DNA synthesis, phosphorylation and cell cycle, in correspondence with the biological functions detailed above.

3.2 Significant terms *vs.* term frequencies

Term frequencies are not good indicators of their relevance, since general terms such *as the, it, and* or other terms of general biological meaning, such as *cell* or *protein*, will always appear at high frequency.

Some terms that appear at a relatively low frequency have considerably significant Z-scores (e.g. *minichromosome maintenance* with frequency 16% and Z-score 2.84). The terms of relatively low frequency have two origins: a) the number of abstracts referring to a given function may be comparatively small, as most articles linked to the gene cluster will address other possible functional aspects related to the cluster, and b) it is possible that the function described by the term will not be present in all proteins of the cluster, a situation that will be more frequent in the less well-defined clusters. We therefore use their Z-score as a comparative value, directly related to the significance of the terms for the different gene clusters. In this case, terms such as *mcm*, *DNA synthesis*, *s-phase* and *cell cycle* achieve high Z-scores and are selected by the system (examples are shown in Table 2).

Table 2: Frequencies and Z-scores of some *terms* from cluster J

Significance	Terms	Frequency	Z-score
Minichromosome maintenance	mcm	0.40	2.84
	Minichromosome maintenance	0.16	2.84
DNA synthesis	Licensing factor	0.07	2.85
	DNA synthesis	0.13	1.96
	S phase	0.24	2.51
Phosphorylation	dbf2	0.19	2.85
	Protein kinase	0.18	2.55
Cell cycle	cdc54	0.12	2.85
	Cell cycle	0.54	2.06
Non-specific (biological)	DNA	0.70	2.49
	Antibody	0.18	2.45
	Protein family	0.11	2.71
	Schizosaccharomyces pombe	0.17	1.70
Non-biological	Family	0.44	2.44
	Apparently	0.12	2.23
	Associated	0.22	2.06
	Depends on	0.05	2.30
	License	0.13	2.85

3.3 Significant terms and gene clustering levels

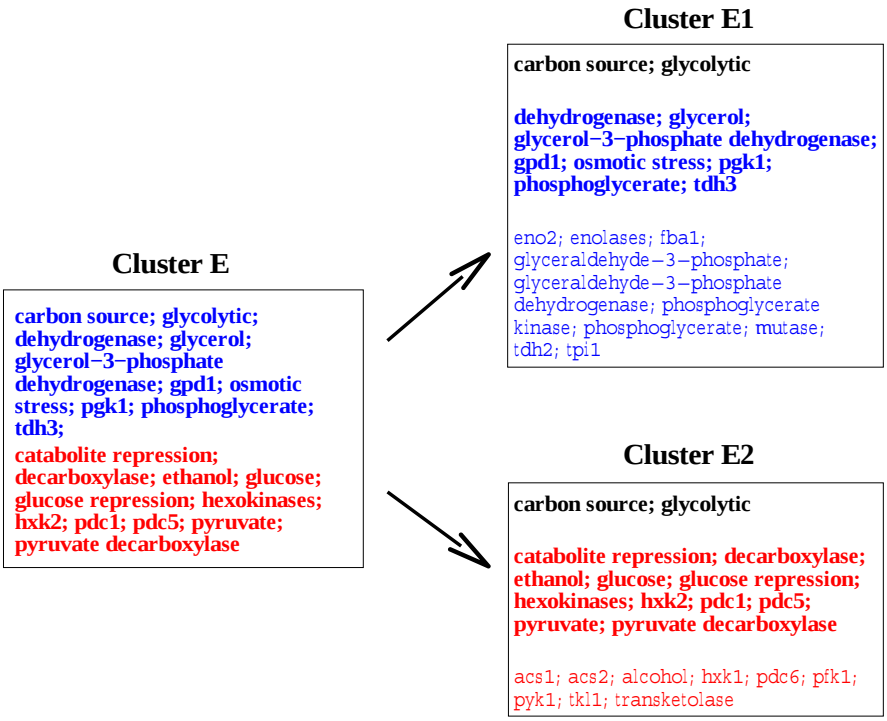


Figure 4: Selected significant terms for cluster E and the derived sub-clusters. Clustering is taken from Eisen *et al.* (1998). Colors indicate the behavior of the terms. The ones in black are general terms for the entire cluster, since they appear both in the root of the classification (initial cluster) and in the derived subclusters. Other terms in blue, red, and highlighted in bold letters correspond to terms that, even if they appear in the initial root cluster, are more specific to some of the sub-clusters. The rest of the terms in italics are specific to the subclusters and do not contain general information for the E cluster (taken from Blaschke 2001).

If a cluster is hierarchically divided into smaller clusters it can be expected that the terms are more general at the higher levels of gene clustering and more specific on a lower level where more similar expression profiles can be found. An example of cluster E (Figure 4, and Eisen *et al.* 1998) can be used to illustrate this point. It is composed mainly of genes related with glycolysis, as detected by the presence of general terms such as carbon source and glycolytic. The further split of the cluster into sub-clusters of more similar expression patterns is clearly correlated with the appearance of terms specific to the two sub-clusters. One of these sub-clusters is better related to the term glycerol whereas the other is better described by terms such as ethanol and pyruvate. This example demonstrates a general trend toward the co-evolution of the similarity of gene expression patterns and the significance of associated terms. Both expression patterns and associated terms became more specific and detailed throughout the clustering process, facilitating the discovery of hidden biological patterns. The questions that will be posed by this type of analysis could include the following: are the differences between glycolytic enzymes, discussed above, related to a possible biochemical origin of the differences in gene expression patterns?

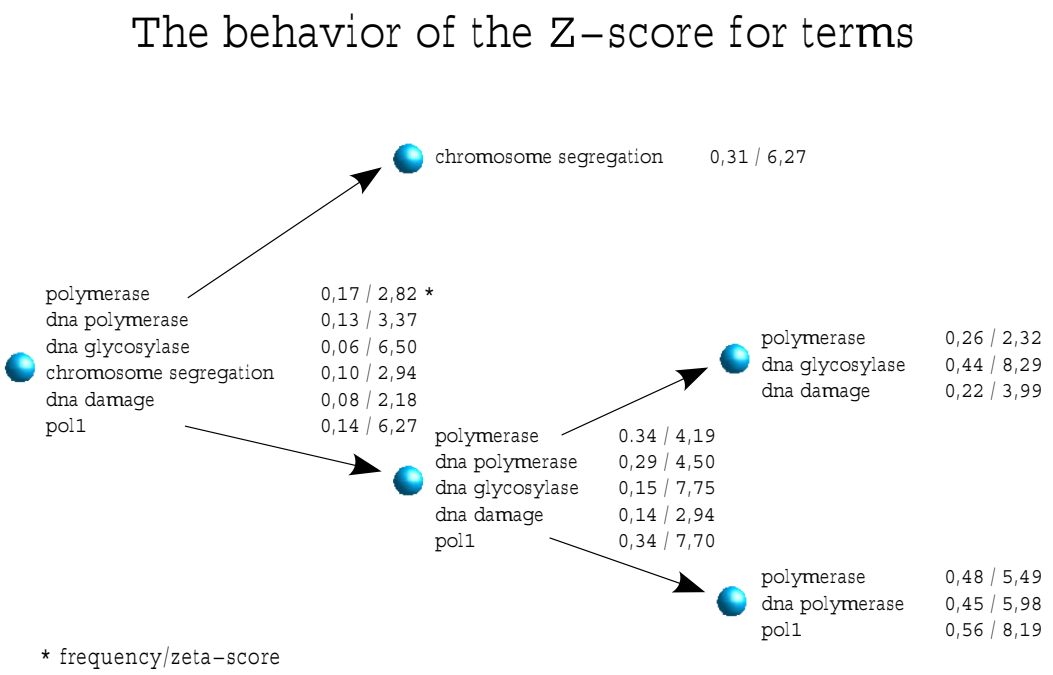


Figure 5: The terms extracted for the groups shown here distribute according to their biological function over the different levels. It is the example of DNA in mitosis where structural and functional aspects are mixed in the root cluster and are then separated into the structural and functional components (taken from Oliveros 2000).

To continue on this line we used a clustering algorithm specifically designed for the analysis of DNA expression arrays (Herrero 2001) to follow this phenomenon over more clustering levels. Figure 5 shows a part of the clustering tree. Two observations can be made. First, in each level the extracted terms separate and the groups get more specific (in this overview many terms are omitted and we focus on a few terms to show the concept of our observation). The cluster on the left has to do with a general aspect of the mitosis, functional and structural part of the DNA. In the next level the functional and structural parts separate, the *chromosome segregation* goes to one side and terms related with DNA replication to the other. Then this group is separated further into DNA polymerization (*pol1*) and DNA repair (*DNA glycosylase, DNA damage*). This shows how the functions of the genes in a group get more and more similar and specific the smaller the groups and the more similar the expression patterns are. The second observation concerns the Z-scores for the terms. In general they grow from one level to the other (meaning that they get more significant), but the score for *polymerase* that is present in all the groups drops at one point. It is present in two groups of the same level and seems to be related with both but not equally. This characteristic can be used for the subsequent functional analysis of the clusters and give the user a dynamic view of how functions are related to the genes at different levels of clustering.

4 Discussion

We propose an application of information extraction techniques for the analysis of expression array data. The increasing complexity of the biological approaches requires the analysis of large

collections of data, such as the expression of thousands of genes in hundreds of conditions that will require development of new methodologies able to organize the information and facilitate the analysis by expert users. The GEISHA system is designed to suggest common functions for the expressed genes by extracting the terms that are differentially represented in large sets of Medline abstracts associated with the distinct gene clusters.

We analyzed the results qualitatively by detailed comparison of automatic and human expert provided annotations. We believe that a quantitative analysis is currently infeasible at least if the evaluation is referred to the biological implications of the extracted information.

Our analysis showed how the information contained in the significant terms was of sufficient biological relevance. In the gene expression experiments analyzed, the systems provided information that would certainly facilitate biological interpretation by human experts, with the obvious advantage of obtaining this information consistently and automatically.

4.1 Coverage of the clusters by the related terms

GEISHA evaluates the significance of the terms associated to a cluster by comparing their frequency with the frequencies of the abstracts containing the terms in the other clusters. The frequencies themselves represent poorly how well the terms cover the functions of the cluster, as frequency does not measure directly whether the terms have a general meaning for the cluster or are related only to a subgroup of genes. For example, a term found at low frequency may correspond to less important terms that would seldom be present in the corresponding abstracts, or to an important term associated to only a small fraction of the genes. In the future we consider providing more detailed information on how terms are related only with subgroups of genes or with the whole cluster.

The terms extracted for a cluster depend on the similarity of their expression profiles. For two examples (glycolysis and DNA in mitosis) we showed that the keywords get more specific and change their significance from one level to the other. Our experience is that they normally get more significant in smaller groups with more similar expression profiles (data not shown), but the exceptions are very interesting and may be used to point the user to inconsistencies or to new biological findings (terms with low significance in a group with very similar expression patterns mean that the information for these genes in the literature is very inhomogeneous and the high similarity of their expression patterns may be a hint to a relation that was not known before).

4.2 Integration with other tools

We have shown that the information obtained by analyzing Medline abstracts can be better understood as complementary to the information provided by different sequence databases, producing a reinforcement of the possible functional annotations. In the future, it would be necessary to incorporate other sources of information, such as the full text of articles, e.g., electronic collections of publications (PubMedCentral 2001; E-BioSci 2001), or annotated data from previous expression array and interaction data derived from different high throughput experiments.

It may be especially interesting to explore the integration with other types of analysis of the text corpus; particularly promising is the inverse analysis based on clustering articles by their composition of keywords (Renner 1999).

Acknoledgements

C. Blaschke implemented the first version of GEISHA, took part in the analysis of the results and prepared the manuscript. JC. Oliveros continued the development of the software and made most of the biological interpretations of the results. L. Cornide implemented the commercial version for ALMA Bioinformatics. A. Valencia developed the initial idea to GEISHA and organized and supervised the work. We are grateful to J. Dopazo and H. Herrero from the CNIO, Madrid to make the clustering algorithm SOTA available to us and support our work in a significant way. Finally we want to thank D. Clark from ALMA Bioinformatics to check the manuscript for language errors and the members of the Protein Design Group for the continuous feed-back on this project. This work was supported in part by TMR grants from the EU.

References

Alizadeh AA, Eisen MB *et al.* (30) (2000) *Distinct types of diffuse large B-cell lymphona identified by gene expression profiling.* Nature 403: 503-511.

Andrade MA and Valencia A (1998) *Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families.* Bioinformatics 14: 600-607.

Bairoch A and Apweiler R (1997) *The SWISS-PROT protein sequence data bank and its supplement TREMBL.* Nucl Acids Res 25: 31-36.

Bassett DE, Eisen MB and Boguski MS (1999). *Gene expression informatics - it's all in your mine.* Nature Genetics Suppl 21: 51-55.

Blaschke C, Oliveros JC and Valencia A (2001). *Mining functional information associated to expression arrays.* Funct Integr Genomics 4, 256-268.

Cho RJ, Campbell MJ ,Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ and Davis RW (1998) *A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle.* Mol Cell 2: 65-73.

Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO and Herskowitz I (1998) *The Transcriptional Program of Sporulation in Budding Yeast.* Science 282: 699-705.

DeRisi JL, Iyer VR and Brown PO (1997) *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science 278: 680-686.

Dr Felix's Free MEDLINE Page (2000) *http://www.beaker.iupui.edu/drfelix/*

E-Bioscience. The electronic publication initiative at EMBO. http://www.embo.org/E_Pub_pages.html

Eisen MB, Spellman PT, Brown PO and Botstein D (1998) *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci USA 95: 14863-14868.

Ermolaeva O, Rastogi M, Pruitt KD, Shuler GD, Bittner ML, Chen Y, Simon R, Meltzer P, Trent JM and Boguski MS (1998) *Data management and anlysis for gene expression arrays.* Nat Gen 20: 19-23.

GEML (Gene Expression Markup Language) at the web site of the Object Management Group for Gene Expression Data: http://www.geml.org/omg.htm

Herrero J, Valencia A and Dopazo J (2001) *A hierarchical unsupervised growing neural network for clustering gene expression patterns.* Bioinformatics 17, 126-136.

Hodges PE, McKee AHZ, Davis BP, Payne WE and Garrels JI (1999) *Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data.* Nucl Acids Res 27: 69-73.

Holstege FCP, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ,Green MR, Golub TR, Lander ES and Young RA (1998) *Dissecting the Regulatory Circuitry of a Eukaryotic Genome.* Cell 95: 717-728.

ImaGene (1999) *ImaGene^{TM}-microarray image analysis software.* BioDiscovery Inc., Los Angeles, CA.

Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J, Boguski MS, Lashkari D, Shalon D, Botstein D and Brown PO (1999) *The trancriptional program in response of human fibroblasts to serum.* Science 283: 83-87.

Jennings EG and Young RA (1999) *Genome expression on the World Wide Web.* TIG 15: 202-203.

Jenssen TK, Lægreid A, Komorowski J and Hovig E (2001). "A literature network of human genes for high-throughput analysis of gene expression". Nature Genetics 28, 21-28.

Liao B, Hale W, Epstein CB, Butow RA and Garner HR (2000) *MAD: a suite of tools for microarray data management and processing.* Bioinformatics 16: 946-947.

Lockhart DJ and Winzeler EA (2000) *Genomics, gene expression and DNA arrays.* Nature 405: 827-836.

Mann PS (1995) *Introductory Statistics.* 2nd ed., 122-124. John Wiley and Sons. New York.

MEDLINE (2001) *http://www.ncbi.nlm.nih.gov/pubmed/ or http://www.nlm.nih.gov/Entrez/medline.html*

Oliveros JC, Blaschke C, Herrero J, Dopazo J and Valencia A (2000). *Expression profiles and biological function.* Genome Informatics Series 11, 106-117.

Proteome Databases (2001). *http://www.proteome.com/databases/index.html*

PubMed Central. A digital archive of life sciences literature managed by the National Center for Biotechnology Information (NCBI). http://www.pubmedcentral.nih.gov

Quackenbush J (2001) *Computational analysis of microarray data.* Nature Reviews Genetics 2, 418-427.

Rebhan M, Chalifa-Caspi V, Prilusky J and Lancet D (1997). *GeneCards: encyclopedia for genes, proteins and diseases.* Weizmann Institute of Science, Bioinformatics Unit and Genome Center.

Renner A and Aszodi A (1999) *High-throughput functional annotation of novel gene products using document clustering.* Pacific Symposium on Biocomputing 2000, 54-65.

Richmond CS, Glasner JD, Mau R, Jin H and Blattner FR (1999) *Genome-wide expression profiling in Escherichia coli K-12.* Nucl Acids Res 27: 3821-3835.

SilverPlatter electronic information provider (2000) *http://www.silverplatter.com/*

Shatkay H, Edwards S, Wilbur WJ and Boguski M (2000) *Genes, Themes, and Microarrays. Using Information Retrieval for Large-Scale Gene Analysis.* ISMB2000, 317-328.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D and Futcher B (1998) *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by nicroarray hybridization.* Mol Bio Cell 9: 3273-3297.

SWISS-PROT(2001) *http://www.expasy.ch/sprot and http://www.ebi.ac.uk/swissprot/*

Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L and Weinstein JN (1999) *MedMiner: An internet text-mining tool for biomedical information, with application to gene expression profiling.* BioTechniques 27, 1210-1217.

Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL and Somogyi R (1998) *Large-scale temporal gene expression mapping of central nervous system development.* Proc Natl Acad Sci USA 95: 334-339.

Wodicka L, Dong H, Mittmann M, Ho MH and Lockhart DJ (1997) *Genome-wide expression monitoring in Saccharomyces cerevisiae.* Nature Biotechnology 15: 1359-1367.

'Deep' Information Extraction from Biomedical Documents in the MEDSYNDIKATE System

Udo Hahn [a] Stefan Schulz [a,b] Martin Romacker [a]
[a] Text Knowledge Engineering Lab
Freiburg University, D-79098 Freiburg, Germany
http://www.coling.uni-freiburg.de

[b] Department of Medical Informatics
Freiburg University Hospital, D-79104 Freiburg, Germany
http://www.imbi.uni-freiburg.de/medinf

Abstract

MEDSYNDIKATE is a natural language processor for automatically harvesting knowledge from medical finding reports. The content of these documents is transferred to formal representation structures which constitute a corresponding text knowledge base. The system architecture integrates requirements from the analysis of single sentences, as well as those of referentially linked sentences forming cohesive texts. The strong demands MEDSYNDIKATE poses to the availability of expressive knowledge sources are accounted for by two alternative approaches to (semi)automatic ontology engineering. We also present data for the knowledge extraction performance of MEDSYNDIKATE for three major syntactic patterns in medical documents.

1 Introduction

The application of methods from the field of natural language processing to biological data has long been restricted to the parsing of molecular structures such as DNA (Searls, 1995; Leung et al., 2001). More recently, however, efforts have also been directed to capturing content from biological documents (research reports, journal articles, etc.), either dealing with restricted information extraction problems such as name recognition for proteins or gene products (Fukuda et al., 1998; Collier et al., 2000; Ono et al., 2001), or more sophisticated ones which aim at the acquisition of knowledge relating to protein or enzyme interactions, molecular binding behavior, etc. (Craven and Kumlien, 1999; Blaschke et al., 1999; Humphreys et al., 2000; Rindflesch et al., 2000).

Current information extraction (IE) systems (for a survey, cf. Cowie and Lehnert (1996)), however, suffer from various weaknesses. First, their range of understanding is bounded by rather limited domain knowledge. The templates these systems are supplied with allow only factual information about particular, a priori chosen entities (cell type, virus type, protein group, etc.) to be assembled from the analyzed documents. Also, these knowledge sources are considered to be entirely static. Accordingly, when the focus of interest of a user shifts to (facets of) a topic not considered so far, new templates must be supplied or existing ones must be updated manually. In any case, for a modified set of templates the analysis has to be rerun for the entire document collection. Templates also provide either no or severely limited inferencing capabilities to reason about the template fillers – hence, their understanding depth is low. Finally, the potential of IE systems for dealing with textual phenomena is rather weak, if it is available at all. Reference relations spanning over several sentences, however, may cause invalid knowledge base structures to emerge so that incorrect information may be retrieved or inferred.

With the SYNDIKATE system family, we are addressing these shortcomings and aim at a more sophisticated level of knowledge acquisition from real-world texts. The source documents we deal

with are currently taken from two domains, *viz.* test reports from the information technology domain for the itSynDiKATe system (Hahn and Romacker, 2000), and medical finding reports, the framework of the medSynDiKATe system (Hahn et al., 1999b). medSynDiKATe is designed to acquire from each input text a maximum number of simple facts ("The *findings correspond to* an *adenocarcinoma.*"), complex propositions ("*All* mucosa layers show an inflammatory infiltration that mainly consists of lymphocytes."), and evaluative assertions ("The findings correspond to a *severe chronical* gastritis."). Hence, our primary goal is to extract conceptually deeper and inferentially richer forms of relational information than that found by state-of-the-art IE systems. Also, rather than restricting natural language processing intentionally to few templates, we here present an open system architecture for knowledge extraction where text understanding is constrained only by the unpredictable limits of available knowledge sources, the domain ontology, in particular.

To achieve this goal, several requirements with respect to language processing proper have to be fulfilled. As most of the IE systems, we require our parser to be robust to underspecification and ill-formed input (cf. the protocols in Hahn et al. (2000)). Unlike almost all of them, our parsing system is particularly sensitive to the treatment of textual reference relations as established by various forms of anaphora (Strube and Hahn, 1999). Furthermore, since SynDiKATe systems rely on a knowledge-rich infrastructure, particular care is taken to provide expressive knowledge repositories on a larger scale. We are currently exploring two approaches. First, we automatically enhance the set of already given knowledge templates through incremental concept learning routines (Hahn and Schnattinger, 1998). Our second approach makes use of the large body of knowledge that has already been assembled in biomedical taxonomies and terminologies (e.g., the UMLS). That knowledge is automatically transformed into a description logics format and, after interactive debugging and refinement, integrated into a large-scale medical knowledge base (Schulz and Hahn, 2000). Besides these engineering issues, we were challenged by the need to properly account for part-whole relations and partonomic reasoning in the biomedical domain. We have developed a parsimonious reasoning model for dealing with concept taxonomies and partonomies (Hahn et al., 1999c) that has recently been extended to mereotopological relations (Schulz and Hahn, 2001) as well.

2 System Architecture

In the following, major design issues for medSynDiKATe are discussed, with focus on the distinction between sentence-level and text-level analysis. We will then turn to two alternative ontology engineering methodologies satisfying the need for the (semi)automatic supply of large amounts of background knowledge. The overall architecture of SynDiKATe is summarized in Figure 1. The general task of any SynDiKATe system consists of mapping each incoming text, T_i, into a corresponding *text knowledge base*, TKB_i, which contains a formal representation of T_i's content. This knowledge will be exploited by various information services, such as inferentially supported fact retrieval.

2.1 Sentence-Level Understanding

Grammatical knowledge for syntactic analysis resides in a fully lexicalized dependency grammar (cf. Hahn et al. (1994) for details), we refer to as *Lexicon* in Figure 1. Basic word forms (lexemes) constitute the leaf nodes of the lexicon tree, while grammatical generalizations from lexemes appear as lexeme class specifications at different levels of abstraction (e.g., nouns and pronouns inherit syntactic properties of the more general lexeme class nominal). The *Generic Lexicon* in Figure 1 contains entries which are domain-independent (such as *move*, *with*, or *month*), while domain-specific extensions are kept in specialized lexicons serving the needs of particular subdomains, e.g., IT (*notebook*, *hard disk*, etc.) or medicine (*adenocarcinoma*, *gastric mucosa*, etc.).

Conceptual knowledge is expressed in a Kl-One-like representation language (cf. Hahn et al. (1999b) for details). These languages support the definition of complex concept descriptions by means of conceptual roles and corresponding role filler constraints which introduce type

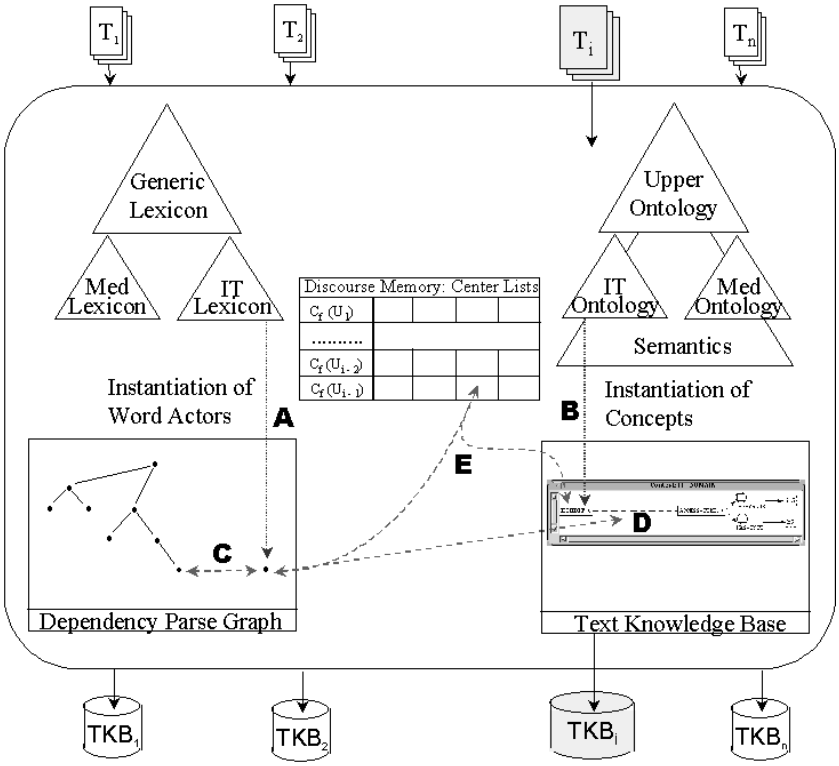


Figure 1: System Architecture of SynDiKATe

restrictions on possible fillers. Taxonomic reasoning can be defined as being primitive (following explicit links), or it can be realized by letting a classifier engine compute subsumption relations between complex conceptual descriptions. A distinction is made between concept classes (types) and instances (representing concrete real-world entities). Most lexemes (except, e.g., pronouns, prepositions) are directly associated with one (or, in case of polysemy, several) concept type(s). Accordingly, when a new lexical item is read from the input text, a dedicated process (word actor) is created for lexical parsing (step A in Figure 1), together with an instance of the lexeme's concept type (step B). Each word actor then negotiates dependency relations by taking syntactic constraints from the already generated dependency tree into account (step C), as well as conceptual constraints supplied by the associated instance in the domain knowledge (step D) (Hahn et al., 2000). As with the *Lexicon*, the ontologies we provide are split up between one that serves all applications, the *Upper Ontology*, while specialized ontologies account for the conceptual structure of particular domains, e.g., information technology (Notebook, Hard-Disk, etc.), or medicine (Adenocarcinoma, Gastric-Mucosa, etc.).

Semantic knowledge is concerned with determining relations between instances of concept classes based on the interpretation of so-called minimal *semantically interpretable subgraphs* of the dependency graph (Romacker et al., 1999). Such a subgraph is bounded by two content words (nouns, verbs, adjectives) which may be directly linked by a single dependency relation or indirectly by a sequence of dependency relations linking non-content words only (e.g., prepositions, auxiliaries). Hence, a conceptual relation may either be constrained by dependency relations (e.g., the *subject:* relation may only be interpreted conceptually in terms of agent or patient roles), by intervening non-content words (e.g., some prepositions impose special role constraints, such as *"with"* does in terms of has-part or instrument roles), or it may only be constrained by conceptual compatibility between the concepts involved (e.g., for genitives). The specification of semantic knowledge shares many commonalities with domain knowledge. Hence, the overlap in Figure 1.

2.2 Text-Level Understanding

The proper analysis of text phenomena prevents inadequate text knowledge representation structures to emerge in the course of sentence-centered analysis (Hahn et al., 1999a). Consider the following text fragment:

(1) Der Befund entspricht einem hochdifferenzierten *Adenokarzinom*.
(The findings correspond to a highly differentiated *adenocarcinoma*.)

(2) *Der Tumor* hat einen Durchmesser von 2 cm.
(*The tumor* has a diameter of 2 cm.)

With purely sentence-oriented analyses, *invalid* knowledge bases are likely to emerge, when each entity which has a different denotation at the text surface is treated as a formally distinct item at the symbol level of knowledge representation, although different denotations refer literally to the same conceptual entity. This is the case for *nominal anaphora*, an example of which is given by the reference relation between the noun phrase *"Der Tumor" (the tumor)* in Sentence (2) and *"Adenokarzinom" (adenocarcinoma)* in Sentence (1). A false referential description appears in Figure 2, where Tumor.2-05 is introduced as a new representational entity, whereas Figure 3 depicts the adequate, intended meaning at the conceptual representation level, *viz.* maintaining Adenocarcinoma.6-04 as the proper referent.

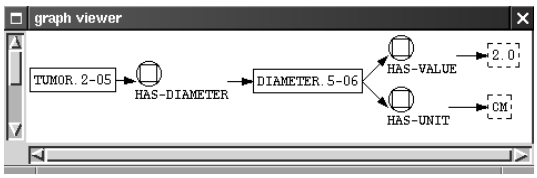


Figure 2: Unresolved Nominal Anaphora

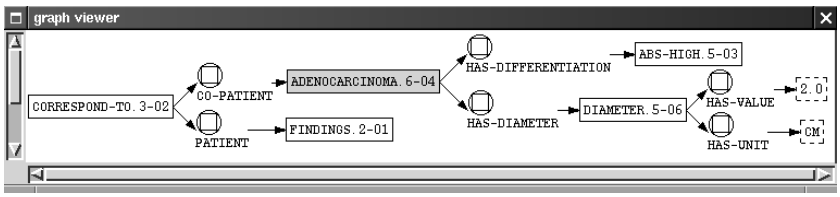


Figure 3: Resolved Nominal Anaphora

The methodological framework for tracking such reference relations at the text level is provided by *center* lists (Strube and Hahn, 1999) (cf. step E in Figure 1). The ordering of their elements indicates that the most highly ranked element is the most likely antecedent of an anaphoric expression in the subsequent utterance, while the remaining elements are ordered according to decreasing preference for establishing referential links.

S_1	[Findings.2-01: Befund, Adenocarcinoma.6-04: Adenokarzinom]
S_2	[Adenocarcinoma.6-04: Tumor, Diameter.5-06: Durchmesser, CM: cm]

Table 1: Center Lists for Sentences (1) and (2)

In Table 1, the tuple notation takes the conceptual correlate of each noun in the text knowledge base in the first place, while the lexical surface form appears in second place. Using the center list of Sentence (1) for the interpretation of Sentence (2) results in a series of queries whether Findings is conceptually more special than Tumor (answer: No) or Adenocarcinoma is more special than Tumor (answer: Yes). As the second center list item for S_1 fulfils all required constraints (mainly the one that Adenocarcinoma is-a Tumor), in the conceptual representation structure of Sentence (2), Tumor.2-05, the literal instance (cf. Figure 2), is replaced by Adenocarcinoma.6-04, the referentially valid identifier (cf. Figure 3). As a consequence, instead of having two unlinked

sentence graphs for Sentence (1) and (2) (e.g., cf. Figure 2) the reference resolution for nominal anaphora leads to joining them in a single coherent and valid text knowledge graph in Figure 3.

Given a fact retrieval application, the validity of text knowledge bases becomes a crucial issue. Disregarding textual phenomena will cause dysfunctional system behavior in terms of incorrect answers. This can be illustrated by a query `Q` such as

```
Q :  (retrieve ?x (Tumor ?x))
A-:  (Tumor.2-05, Adenocarcinoma.6-04)
A+:  (Adenocarcinoma.6-04)
```

which triggers a search for all instances in the text knowledge base that are of type TUMOR. Given an invalid knowledge base (cf. Figure 2), the incorrect answer (A-) contains two entities, *viz.* TUMOR.2-05 and ADENOCARCINOMA.6-04, since both are in the extension of the concept TUMOR. If, however, a valid text knowledge base such as the one in Figure 3 is given, only the correct answer, ADENOCARCINOMA.6-04, is inferred (A+).

2.3 ONTOLOGY ENGINEERING

MEDSYNDIKATE requires a knowledge-rich infrastructure both in terms of grammar and domain knowledge, which can hardly be maintained by human efforts alone. Rather a significant amount of knowledge should be generated automatically. For SYNDIKATE systems, we have chosen a dual strategy. One focuses on the incremental learning of new concepts while understanding the texts, the other is based on the reuse of available comprehensive (though semantically weak) knowledge sources.

Concept Learning from Text. Extending a given core ontology by new concepts as a by-product of the text understanding process builds on two different sources of evidence — the already given domain knowledge, and the grammatical constructions in which unknown lexical items occur in the source document. The parser yields information from the grammatical constructions in which lexical items occur in terms of the labellings in the dependency graph. The kinds of syntactic constructions in which unknown words appear are recorded and later assessed relative to the credit they lend to a particular hypothesis. Typical linguistic indicators that can be exploited for taxonomic integration are, e.g., appositions *('the symptom @A@'*, with '@A@' denoting the unknown word) or exemplification phrases *('symptoms like @A@')*, These constructions almost unequivocally determine '@A@' when considered as a medical concept to denote an instance of a SYMPTOM.

The conceptual interpretation of parse trees involving unknown words in the text knowledge base leads to the derivation of concept hypotheses, which are further enriched by conceptual annotations. These reflect structural patterns of consistency, mutual justification, analogy, etc. relative to already available concept descriptions in the ontology or other concept hypotheses. Grammatical and conceptual evidence of this kind, in particular their predictive "goodness" for the learning task, are represented by corresponding sets of linguistic and conceptual quality labels. Multiple concept hypotheses for each unknown lexical item are organized in terms of hypothesis spaces, each of which holds alternative or further specialized conceptual readings. An inference engine coupled with the classifier, the so-called quality machine, estimates the overall credibility of single concept hypotheses by taking the available set of quality labels for each hypothesis into account (cf. Hahn and Schnattinger (1998) for details).

Reengineering Medical Terminologies. The second approach makes use of the large body of knowledge that has already been assembled in comprehensive medical terminologies such as the UMLS (NLM, 2001). The knowledge they contain, however, cannot be fed directly to MEDSYNDIKATE, because it is characterized by inconsistencies, circular definitions, insufficient depth, gaps, etc., and the lack of an inference engine.

The methodology for reusing weak medical knowledge consists of four steps (Schulz and Hahn, 2000). First, we create automatically KL-ONE-style logical expressions by feeding a generator with data directly from the UMLS, i.e., the concepts and the semantic links between concept pairs. In a second step, the imported concepts, already in a logical format, are submitted to the classifier of the knowledge representation system (in our case, LOOM) in order to check

whether the terminological definitions are consistent and non-circular. For those elements which are inconsistent, their validity is restituted and definitional circles are removed manually by a medical domain expert. In the final step the knowledge base which has emerged so far is manually rectified and refined (e.g., by checking the adequacy of taxonomic and partonomic hierarchies).

3 Evaluating Knowledge Extraction Performance

3.1 Evaluation Framework

In quantitative terms, SynDiKATe is neither a toy system nor a monster. The *Generic Lexicon* currently includes 5,000 entries, while the *MED Lexicon* contributes 3,000 entries. Similarly, the *Upper Ontology* contains 1,500 concepts and roles, to which the *MED Ontology* adds 2,500 concepts and roles. However, recent experiments with reengineering the UMLS have resulted in a very large medical knowledge base with 164,000 concepts and 76,000 relations (Schulz and Hahn, 2000) that is currently under validation.

We extracted the text collection from the hospital information system of the University Hospital in Freiburg (Germany). All finding reports in histopathology from the first quarter of 1999 were initially included, altogether 4,973 documents. However, for the time being medSynDiKATe covers especially the subdomain of gastro-intestinal diseases. Thus, 752 texts out of these 4,973 were extracted semi-automatically in order to guarantee a sufficient coverage of domain knowledge. From this collection, a random sample of 90 texts was taken and divided into two sets. 60 of them served as the training set which was used for parameter tuning of the system. The remaining 30 texts were then used to measure the performance of the medSynDiKATe system with unseen data. The configuration of the system was frozen prior to analyzing the test set.

In the empirical study proper, three basic settings of dependency graphs were evaluated, *viz.* ones containing genitives, prepositional phrases, as well as constructions including modal verbs or auxiliaries. Genitives and prepositional phrases relate fundamental biomedical concepts via associated roles at the conceptual level. Modal and auxiliary verbs create a complex syntactic environment for the interpretation of verbs, and, hence, the conceptual representation of medical processes and events. For each instance of these configurations semantic interpretations were automatically computed the result of which was judged for accuracy by two skilled raters.

Still, the way how a (gold) standard for semantic interpretation can be set up is an issue of hot debates (Zweigenbaum et al., 1997). In fact, conceptually annotated medical text corpora do not exist at all, at least for the German language. At this level, the ontology we have developed eases judgements, since it is based on a fine-grained relation hierarchy with clear sortal restrictions for role fillers. In anatomy, e.g., we use relations such as anatomical-part-of, which is itself a subrelation of physical-part-of and part-of, and specialize it in order to account for subtle part-of relationships. A very specific relation such as anatomical-part-of-mucosa refers to a precise subset of entities to be related by the interpretation process. Therefore, relating Brain to Mucosa by anatomical-part-of-mucosa obviously would be considered as incorrect, whereas relating Lamina-Propria-Mucosae would be considered a reasonable interpretation.

3.2 Quantitative Analysis

The following tables contain data for both the training and the test set indicating the quality of knowledge extraction as obtained for the three different syntactic settings (for additional data, cf. Romacker and Hahn (2000)). Besides providing data for recall and precision, the tables are divided into two assessment layers: *"without interpretation"* means that the system was not able to produce an interpretation because of specification gaps, i.e., at least one of the two content words in a minimal dependency graph under consideration was not specified. Note that even for the training set which was intended to generate optimal results we were unable to formulate reasonable and generally valid concept definitions for some of the content words we encountered (e.g., for fuzzy expressions of locations: *"In der Tiefe der Schleimhaut"* (*"In the depth of the mucosa"*)). The second group *"with interpretation"* is divided into four categories. The label *correct (non-ambiguous)* qualifies, if just a single and correct conceptual relation was computed by

the semantic interpretation process. However, if the result was correct but yielded more than one conceptual relation, the label *correct (ambiguous)* was assigned. An interpretation was considered *incorrect* when the conceptual relation was inappropriate. Finally, *NIL* was used to indicate that an interpretation was performed (both concepts for the content words were specified) but no conceptual relation could be computed.

Genitives. In the medical domain, as indicated by Table 2 the recall and precision values for the interpretation of genitives are very encouraging both for the training set (92% and 93%) and the test set (93% and 93%), respectively.[1] However, since genitives, in general, provide no additional constraints how the conceptual correlates of the two content words involved can be related, the number of ambiguous interpretations amounts to 13% and 36%, respectively.

	Training Set	Test Set
Recall	92%	93%
Precision	93%	93%
# occurrences ...	168	91
... **with interpretation**	158 (94%)	86 (95%)
[confidence intervals]	[89%-97%]	[90%-98%]
..... correct (non-ambiguous)	125.5 (75%)	48.5 (53%)
..... correct (ambiguous)	22 (13%)	33 (36%)
..... incorrect	6.5	3.5
..... NIL	4	1
... **without interpretation**	10 (6%)	5 (5%)

Table 2: Evaluation of Genitives

Auxiliaries and Modals. Table 3 contains the results for modal verbs or auxiliaries. A semantic interpretation of modal/auxiliary verb complexes relates a content-bearing verb with the conceptual correlate of the syntactic subject. In case of a passive construction the direct-object-to-subject normalization has to be carried out.

	Training Set	Test Set
Recall	94%	80%
Precision	98%	84%
# occurrences ...	131	55
... **with interpretation**	125 (95%)	52 (95%)
[confidence intervals]	[92%-99%]	[84%-99%]
..... correct (non-ambiguous)	122 (93%)	43,5 (79%)
..... correct (ambiguous)	1	0
..... incorrect	0	0,5
..... NIL	2	8 (15%)
... **without interpretation**	6 (5%)	3 (5%)

Table 3: Evaluation of Modal Verbs and Auxiliaries

Recall and precision for the training set are high (94% and 98%, respectively) and, therefore, indicate that semantic interpretation can cope with almost all occurrences given an optimal degree of specification. The values for recall and precision dropped to 80% and 84%, respectively, in the test set. The increase of *NIL* results reveals that the granularity of the underlying domain model is insufficient as far as conceptual relations are concerned. Although the corresponding concepts are modelled, no conceptual relation between them could be determined.

Prepositional phrases (PPs) are crucial for the semantic interpretation of a text, since they introduce a wide variety of conceptual relations, such as spatial, temporal, causal, or instrumental

[1] Confidence intervals for .95 probability are given in square brackets.

ones. The importance of PPs is reflected by their relative frequency. In the training set and the test set, we encountered 1,108 prepositions, which is a little bit less than 10% of the words in both sets (approximately 11,300).[2] Provided also that the preposition's syntactic head and its modifier participate in the interpretation, at the phrase level, more than 25% of the texts' contents is encoded by PPs (certainly, this data also reflects a considerable degree of genre dependency).

	Training Set	Test Set
Recall	85%	85%
Precision	79%	81%
# occurrences ...	562	278
... **with interpretation**	548 (98%)	253 (91%)
[confidence intervals]	[96%-99%]	[86%-93%]
..... correct (non-ambiguous)	401,5 (71%)	167 (60%)
..... correct (ambiguous)	32,5 (6%)	37,5 (13%)
..... incorrect	43 (8%)	30,5 (11%)
..... NIL	71 (13%)	18 (6%)
... **without interpretation**	14 (2%)	25 (9%)

Table 4: Evaluation of Prepositional Phrases

Considering the results for semantic interpretation of PPs (cf. Table 4), the values for recall and precision are almost the same for the training set and the test set. Recall climaxed at 85% for both the training set and the test set, whereas precision reached 79% for the training set and 81% for the test set. Getting almost the same performance for both sets also reveals a stable level of semantic interpretation of PPs.[3]

4 Conclusions

We have introduced medSynDiKATe, a system for harvesting knowledge from biomedical reports. Emphasis was put on the role of various knowledge sources required for 'deep' text understanding. When turning from sentence-level to text-level analysis, we considered representational inadequacies when text phenomena were not properly accounted for and, hence, proposed a solution based on centering mechanisms.

The enormous knowledge requirements posed by our approach can only be reasonably met when knowledge engineering does not rely on human efforts only. Hence, a second major issue we have focused on concerns alternative ways to support knowledge acquisition and guarantee, this way, a reasonable chance for scalability of the system. We made two proposals. The first one deals with an automatic concept learning methodology that is fully embedded in the text understanding process, the other one exploits the vast amounts of medical knowledge assembled in various knowledge repositories such as the UMLS.

We, finally, provided empirical data which characterizes the knowledge extraction performance of medSYNDIKATE on three major syntactic structures, *viz.* genitives, modals and auxiliaries, and prepositional phrases. These reflect, at the linguistic level, fundamental categories of biomedical ontologies — states, processes, and actions.

[2] Only 940 of these 1,108 were included in the empirical analysis, since 168 did not form a minimal subgraph. Phrases like *"zum Teil"* (*"partly"*) map to a single meaning — as evidenced by the English translation correlate — and were therefore excluded.

[3] Corresponding data for an alternative test scenario, knowledge extraction from information technology (IT) test reports, is not as favorable as for the medical domain. For PPs, 70% / 64% recall and 77% / 66% precision were measured for the training set and the test set, respectively. A reasonable argument why we achieved better results in the medical domain than in the IT world might be that in the medical texts a considerably lower degree of linguistic variation is encountered.

References

Blaschke, C., Andrade, M., Ouzounis, C., and Valencia, A. (1999). Automatic extraction of biological information from scientific text: Protein-protein interactions. *Intelligent Systems for Molecular Biology*, 7:60–67.

Collier, N., Nobata, C., and Tsujii, J.-i. (2000). Extracting the names of genes and gene products with a hidden Markov model. In *COLING 2000 – Proceedings of the 18th International Conference on Computational Linguistics*, volume 1, pages 201–207. Saarbrücken, Germany, 31 July - 4 August, 2000. San Francisco, CA: Morgan Kaufmann.

Cowie, J. and Lehnert, W. (1996). Information extraction. *Communications of the ACM*, 39(1):80–91.

Craven, M. and Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In *ISMB'99 – Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, pages 77–86. Heidelberg, Germany, August 6-10, 1999. AAAI Press.

Fukuda, F., Tsunoda, T., Tamura, A., and Takagi, T. (1998). Toward information extraction: Identifying protein names from biological papers. In Altman, R. B., Dunker, A. K., Hunter, L., and Klein, T. E., editors, *PSB 98 – Proceedings of the 3rd Pacific Symposium on Biocomputing*, pages 705–716. Maui, Hawaii, USA, 4-9 January, 1998. Singapore: World Scientific Publishing Co.

Hahn, U., Bröker, N., and Neuhaus, P. (2000). Let's ParseTalk: Message-passing protocols for object-oriented parsing. In Bunt, H. and Nijholt, A., editors, *Advances in Probabilistic and Other Parsing Technologies*, volume 16 of *Text, Speech and Language Technologies*, pages 177–201. Dordrecht, Boston: Kluwer.

Hahn, U. and Romacker, M. (2000). Content management in the SynDiKATe system: How technical documents are automatically transformed to text knowledge bases. *Data & Knowledge Engineering*, 35(2):137–159.

Hahn, U., Romacker, M., and Schulz, S. (1999a). Discourse structures in medical reports – watch out! The generation of referentially coherent and valid text knowledge bases in the medSynDiKATe system. *International Journal of Medical Informatics*, 53(1):1–28.

Hahn, U., Romacker, M., and Schulz, S. (1999b). How knowledge drives understanding: Matching medical ontologies with the needs of medical language processing. *Artificial Intelligence in Medicine*, 15(1):25–51.

Hahn, U., Schacht, S., and Bröker, N. (1994). Concurrent, object-oriented natural language parsing: The ParseTalk model. *International Journal of Human-Computer Studies*, 41(1/2):179–222.

Hahn, U. and Schnattinger, K. (1998). Towards text knowledge engineering. In *AAAI'98/IAAI'98 – Proceedings of the 15th National Conference on Artificial Intelligence & 10th Conference on Innovative Applications of Artificial Intelligence*, pages 524–531. Madison, Wisconsin, July 26-30, 1998. Menlo Park, CA & Cambridge, MA: AAAI Press & MIT Press.

Hahn, U., Schulz, S., and Romacker, M. (1999c). Part-whole reasoning: A case study in medical ontology engineering. *IEEE Intelligent Systems & Their Applications*, 14(5):59–67.

Humphreys, K., Demetriou, G., and Gaizauskas, R. (2000). Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. In Altman, R. B., Dunker, A. K., and Hunter, L., editors, *PSB 2000 – Proceedings of the 5th Pacific Symposium on Biocomputing*, pages 502–513. Honolulu, Hawaii, USA, 4-9 January, 2000. Singapore: World Scientific Publishing Co.

Leung, S.-w., Mellish, C., and Robertson, D. (2001). Basic gene grammars and DNA-ChartParser for language processing of *Escherichia coli* promoter DNA sequences. *Bioinformatics*, 17(3):226–236.

NLM (2001). *Unified Medical Language System*. Bethesda, MD: National Library of Medicine.

Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. (2001). Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161.

Rindflesch, T. C., Rajan, J. V., and Hunter, L. (2000). Extracting molecular binding relationships from biomedical text. In *ANLP 2000 – Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 188–195. Seattle, Washington, USA, April 29 - May 4, 2000. San Francisco, CA: Morgan Kaufmann.

Romacker, M. and Hahn, U. (2000). An empirical assessment of semantic interpretation. In *NAACL 2000 – Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 327–334. Seattle, Washington, USA, April 29 - May 4, 2000. San Francisco, CA: Morgan Kaufmann.

Romacker, M., Markert, K., and Hahn, U. (1999). Lean semantic interpretation. In *IJCAI'99 – Proceedings of the 16th International Joint Conference on Artificial Intelligence*, volume 2, pages 868–875. Stockholm, Sweden, July 31 - August 6, 1999. San Francisco, CA: Morgan Kaufmann.

Schulz, S. and Hahn, U. (2000). Knowledge engineering by large-scale knowledge reuse: Experience from the medical domain. In Cohn, A. G., Giunchiglia, F., and Selman, B., editors, *Principles of Knowledge Representation and Reasoning. Proceedings of the 7th International Conference – KR 2000*, pages 601–610. Breckenridge, Colorado, USA, April 12-15, 2000. San Francisco, CA: Morgan Kaufmann.

Schulz, S. and Hahn, U. (2001). Parts, locations, and holes: Formal reasoning about anatomical structures. In Quaglini, S., Barahona, P., and Andreassen, S., editors, *Artificial Intelligence in Medicine. Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe – AIME 2001*, volume 2101 of *Lecture Notes in Artificial Intelligence*, pages 293–303. Cascais, Portugal, July 1-4, 2001. Berlin: Springer.

Searls, D. B. (1995). String variable grammar: A logic grammar formalism for the biological language of DNA. *Journal of Logic Programming*, 24(1/2):73–102.

Strube, M. and Hahn, U. (1999). Functional centering: Grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344.

Zweigenbaum, P., Bouaud, J., Bachimont, B., Charlet, J., and Boisvieux, J.-F. (1997). Evaluating a normalized conceptual representation produced from natural language patient discharge summaries. In Masys, R., editor, *AMIA'97 – Proceedings of the 1997 AMIA Annual Fall Symposium (formerly SCAMC). The Emergence of 'Internetable' Health Care: Systems that Really Work*, pages 590–594. Nashville, TN, October 25-29, 1997. Philadelphia, PA: Hanley & Belfus.

Information Extraction from Biomedical Text

Jerry R. Hobbs
Artificial Intelligence Center
SRI International
Menlo Park, California 94025
hobbs@ai.sri.com

Information extraction is the process of scanning text for information relevant to some inteest, including extracting entities, relations, and, most challenging, events—or who did what to whom. It requires deeper analysis than key word searches, but its aims fall short of the very hard and long-term problem of text understanding, where we seek to capture *all* the information in a text, along with the speakers' or writer's intention. Information extraction represents a midpoint on this spectrum, where the aim is to capture structured information without sacrificing feasibility. In the last ten years, the technology of information extraction has advanced significantly. It has been applied primarily to domains of economic and military interest. The are now initial efforts to apply it to biomedical text, and the time is ripe for further research.

One of the key ideas in this technology is to separate processing into several stages. in "cascaded finite-state transducers". The earlier stages recognize smaller linguistic objects and work in a largely domain-independent fashion. They use purely linguistic knowledge to recognize that portion of the syntactic structure of the sentence that linguistic methods can determine reliably, requiring little or no modification or augmentation as the system is moved from domain to domain, at least prior to moving to the biomedical domain.

The later stages take these linguistic objects as input and find domain-dependent patterns among them.

Typically there are five levels of processing:

1. Complex Words: This includes the recognition of multiwords and proper names. In biomedicine thiw would include names of chemical compounds.

2. Basic Phrases: Sentences are segmented into noun groups, verb groups, and particles.

3. Complex Phrases: Complex noun groups and complex verb groups are identified.

4. Domain Patterns: The sequence of phrases produced at Level 3 is scanned for patterns of interest to the application, and when they are found, semantic structures are built that encode the information about entities and events contained in the pattern.

5. Merging Structures: Semantic structures from different parts of the text are merged if they provide information about the same entity or event.

As we progress through the five levels, larger segments of text are analyzed and structured.

This decomposition of the natural-language problem into levels is essential to the approach. Many systems have been built to do pattern matching on strings of words. The advances in information extraction have depended crucially on dividing that process into separate levels for recognizing phrases and recognizing patterns. Phrases can be recognized reliably with purely syntactic information, and they provide precisely the elements that are required for stating the patterns of interest.

I will illustrate the levels of processing by describing what is done on the following sentences, from a biomedical abstract.

> gamma-Glutamyl kinase, the first enzyme of the proline biosynthetic pathway, was purified to a homogeneity from an Escherichia coli strain resistant to the proline analog 3,4-dehydroproline. The enzyme had a native molecular weight of 236,000 and was apparently comprised of six identical 40,000-dalton subunits.

In this example, we will assume we are mapping the information into a complex database of pathways, reactions, and chemical compounds, such as the EcoCyc database developed by Peter Karp and his colleagues at SRI International. In this database there are Reaction objects with the attributes ID, Pathway, and Enzyme, among others, and Enzyme objects with the attributes ID, Name, Molecular-Weight, Subunit-Component, and Subunit-Number.

The five phases are as follows:

1. Complex Words: This level of processing identifies multiwords such as "gamma-Glutamyl proline", Escherichia coli", "3,4-dehydroproline", and "molecular weight"

Languages in general are very productive in the construction of short, multiword fixed phrases and proper names employing specialized microgrammars. The biomedical language is especially rich in this regard. This is the level at which they are recognized.

2. Basic Phrases: At Level 2 the first example sentence is segmented into the following phrases:

Enzyme Name:	gamma-Glutamyl kinase
Noun Group:	the first enzyme
Preposition:	of
Noun Group:	the proline biosynthetic pathway
Verb Group:	was purified
Preposition:	to
Noun Group:	homogeneity
Preposition:	from
Noun Group:	an Escherichia coli strain
Adjective Group:	resistant
Preposition:	to
Noun Group:	the proline analog
Noun Group:	3,4-dehydroproline

Noun groups are noun phrases up through the head noun but not including the right modifiers like prepositional phrases and relative clauses. Verb groups are head verbs with their auxilliaries. Adjective phrases are predicate adjectives together with their copulas, if present.

This breakdown of phrases into nominals, verbals, and particles is a linguistic universal. Whereas the precise parts of speech that occur in any language can vary widely, every language has elements that are fundamentally nominal in character, elements that are fundamentally verbal or predicative, and particles or inflectional affixes that encode relations among the other elements.

3. Complex Phrases: At Level 3, complex noun groups and verb groups that can be recognized reliably on the basis of domain-independent, syntactic information are recognized. This includes the attachment of appositives to their head noun group,

> the proline analog 3,4-dehydroproline

and the attachment of "of" prepositional phrases to their head noun groups,

> the first enzyme of the proline biosynthetic pathway.

In the course of recognizing basic and complex phrases, entities and events of domain interest are often recognized, and the structures for these are constructed. In the sample text, an Enzyme structure is constructed for gamma-Glutamyl kinase. Corresponding to the complex noun group "gamma-Glutamyl kinase, the first enzyme of the proline biosynthetic pathway," the following structure are built:

Reaction:

ID:	R1
Pathway:	proline
Enzyme:	E1

Enzyme:

ID:	E1
Name:	gamma-Glutamyl kinase
Molecular-Weight:	–
Subunit-Component:	–
Subunit-Number:	–

In many languages some adjuncts are more tightly bound to their head nouns than others. "Of" prepositional phrases are in this category, as are phrases headed by prepositions that the head noun subcategorizes for. The basic noun group together with these adjuncts constitutes the complex noun group. Complex verb groups are also motivated by considerations of linguistic universality. Many languages have quite elaborate mechanisms for constructing complex verbs. One example in English is the use of control verbs; "to conduct an experiment" means the same as "to experiment". Another example is the verb-particle constructions such as "set up".

4. Clause-LevelDomain Patterns: In the sample text, the domain patterns

<Compound> have <Measure> of <values>
<Compound> comprised of<Compound>

are instantiated in the second sentence. These patterns result in the following Enzyme structures being built:

Enzyme:

ID:	E2
Name:	–
Molecular-Weight:	236,000
Subunit-Component:	–
Subunit-Number:	–

Enzyme:

ID:	E3
Name:	–
Molecular-Weight:	–
Subunit-Component:	E4
Subunit-Number:	6

Enzyme:

ID:	E4
Name:	–
Molecular-Weight:	40,000
Subunit-Component:	–
Subunit-Number:	–

This level corresponds to the basic clause level that characterizes all languages, the level at which in English Subject-Verb-Object (S-V-O) triples occur, and thus again corresponds to a linguistic universal. This is the level at which predicate-argument relations between verbal and nominal elements are expressed in their most basic form.

5. Merging Structures: The first four levels of processing all operate within the bounds of single sentences. The final level of processing operates over the whole discourse. Its task is to see that all the information collected about a single entity or relationship is combined into a unified whole. This is where the problem of coreference is dealt with in this approach.

The three criteria that are taken into account in determining whether two structures can be merged are the internal structure of the noun groups, nearness along some metric, and the consistency, or more generally, the compatibility of the two structures.

In the analysis of the sample text, we have produced four enzyme structures. Three of them are consistent with each other. Hence, they are merged, yielding

Enzyme:	
ID:	E1
Name:	gamma-Glutamyl kinase
Molecular-Weight:	236,000
Subunit-Component:	E4
Subunit-Number:	6

The fourth is inconsistent because of the differing molecular weights and the subunit relation, and hence is not merged with the others.

The finite-state technology has sometimes beencharacterized as *ad hoc* and as *mere* pattern-matching. However, the approach of using a *cascade* of finite-state machines, where each level corresponds to a linguistic natural kind, reflects important universals about language. It was inspired by the remarkable fact that very diverse languages all show the same nominal element - verbal element - particle distinction and the basic phrase - complex phrase distinction. Organizing a system in this way leads to greater portability among domains and to the possibility of easier acquisition of new patterns.

Information extraction is evaluated by two measures—recall and precision. Recall is a measure of completeness, precision of correctness. When you promise to tell the whole truth, you are promising 100% recall. When you promise to tell nothing but the truth, you are promising 100% precision.

In Message Understanding Conference (MUC) evaluations in the 1990s, systems doing name recognition achieved about 95% recall and precision, which is nearly human-level performance, and very much faster. In event recognition the performance plateaued at about 60% recall and precision.

There are several possible reasons for this. Our analysis of our results showed that the process of merging was implicated in a majority of our errors; we need better ways of doing event and relationship coreference. It could be that 60% is how much information texts "wear on their sleeves". Current technology can only extract what is explicit in texts . To get the rest of the information requires inference. A third possibility is that the distribution of linguistic phenomena simply has a very long tail. Handling the most common phenomena gets you to 60% relatively quickly. Getting to 100% then requires handling increasingly rare phenomena. A month's work gets you to 60%. Another year's work gets you to 65%.

This raises the interesting question of what utility there is in a 60% technology. Obviously you would not be happy with a bank statement that is 60% accurate. On the other hand, 60% accuracy in web search would be a distinct improvement. It is best to split this question into two parts—recall and precision.

If you have 60% recall, you are missing 40% of the mentions of relevant information. But there are half a million biomedical articles a year, and keeping up with them requires massive curatorial effort. 60% recall is an improvement if you would otherwise have access to much less. Moreover, recall is measured not on facts but on *mentions* of facts. If there are multiple mentions of some fact, we have multiple op0portunities to capture it.

With 60% precision in a fully automatic system, then 40% of the information in your database will be wrong. You need a human in the loop. This is not necessarily a disaster. A personextracting sparse information from a massive corpus will have a much easier time discarding 40% of the entries than locating and entering 60%. Good tools would help in this as well. In addition it may be that the usage of language in biomedical text is tightly enough constrained that percision will be higher than in the domains that have so far been the focus of efforts in informaiton extraction.

Ontology Driven Information Extraction*

U. Reyle
Institute for Computational Linguistics
University of Stuttgart
Uwe.Reyle@ims.uni-stuttgart.de

J. Šarić
European Media Laboratory
Heidelberg
Jasmin.Saric@eml.villa-bosch.de

Abstract

We describe the linguistic components of a system (GenIE) that automatically extracts information about biochemical pathways from free text sources. We show that the extraction of information must be based on specialized lexica, semantic representations and ontologies. Furthermore a case study is presented that motivates the use of deductive schemata and a systematic relationship between low-level and high-level representations for the interleaving of shallow and deep processing.

Keywords: information extraction, semantics, ontology, architecture, biochemistry

1 Introduction

The analysis and interpretation of biochemical data is crucial for the understanding of biochemical phenomena and the discovery of new concepts. The organisation of the scientific data in a model that captures the concepts and relations involved can facilitate the understanding and reasoning over the data. This is the overall objective of the EML [1] *Data Alive* project. The primary goal of *Data Alive* lies in the development of a database which allows the handling and representation of the complex heterogeneous data about biochemical pathways. The creation of this database system has been done using an ontology for biochemical pathways, which includes concepts ranging from basic biochemistry to genomics.

Regarding the explosive growth of literature in this area the gains of computational linguistics and automatic information extraction technology have to be exploited to feed the databases from free text sources. Conventional extraction technology has been successfully applied to very restricted extraction tasks, like protein-protein interactions. The objective of the GenIE[2] project is more complex. Its aim is to extract information about biochemical pathways and about sequences, structures and functions of genomes and proteins.

We will show that the extraction of information must be based on specialized lexica and parsers as well as semantic representations and ontologies. Only then can information extraction technology successfully be applied to more complex extraction tasks. Obviously the incorporation of these knowledge sources into the extraction system leads to the requirement to combine shallow and efficient NLP techniques - like tagging and chunking as well as template filling - with a deeper semantic interpretation in case more semantic insight is needed. This feature of 'depth if required'

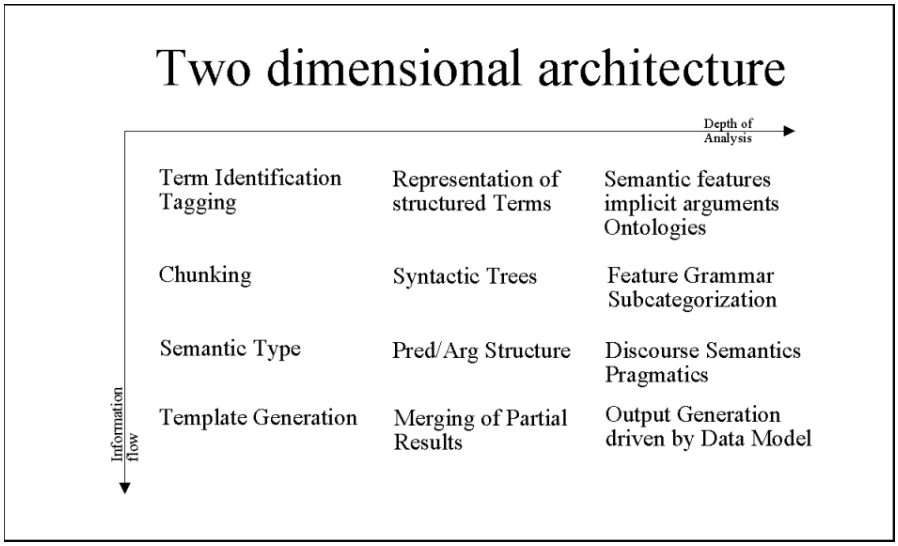
*Part of this work has been supported by the Klaus Tschira Foundation gGmbH, Heidelberg

[1] European Media Laboratory GmbH, Heidelberg

[2] GenIE abbreviates Genome Information Extraction. After a preparatory phase of one year the project started the first of January 01. More information can be found at
http://www.ims.uni-stuttgart.de/projekte/GenIE/main.html

distinguishes our approach to information extraction from text understanding on the one hand and standard information extraction on the other.

In order to fulfill this requirement of interleaving shallow and deep processing there must be a systematic relationship between the representations used in shallow processing on the one hand and the representations contemplated in lingusitic theories one the other hand. This is best achieved by a methodology in which low-level and high-level representations are developed in conjunction and embedded into a comprehensive, computational architecture that guarantees possibilities for information flow as illustrated in the following picture.



Domain specific ontological and semantical knowledge can be exploited at any stage of morphological and syntactic processing, or vice versa. This two-dimensionality leaves sufficient leeway to accommodate combinations of shallow parsing as well as deep semantic analysis, and of full parsing and flat semantics. The two-dimensional architecture should also be shared by the morphological analyzer.

The extraction process itself is mediated by a semantic representation of the text which is then used to fill the data base. This semantic representation will make contextual information explicit and will serve as a basis for disambiguation tasks and contextual resolution. The semantic representations (at any level) are required not to overspecify, i.e. not to force any ambiguity the resolution of which is irrelevant for the extraction task at hand.

In Section 2 we illustrate the "generic" components of GenIE. Section 3 presents a case analysis in which we will factor out the minimal semantic and ontological assumptions that are needed for the extraction task. Deeper semantic relationships are achieved by inference rules. Section 4 illustrates the basic need for a lexicon of biochemical terms and morphemes, and a morphological analyzer for derivationally and compositionally complex expressions. Section 5 presents the domain-specific ontology which comprises a theory of the atomic predicates and relations used in the lexicon.

2 The "generic" components of GenIE

The following example (taken from "The Journal of Biological Chemistry" (Proost et al. (1995))) will illustrate the components of GenIE that had been developed mainly at IMS (University of Stuttgart) and European Media Lab Heidelberg (EML), and have been adapted to the domain of biochemical texts.

> *Phosphofructokinase (EC 2.7.1.11) (PFK) catalyzes the transfer of the gamma-phosphate from MgATP to fructose 6-phosphate (Fru-6P) in the first committed step of glycolysis.*

In a very first step of the analysis preprocessing has to take place. This means the conversion from HTML or PDF to plain text. Then the following steps are performed.

1. **Tokenization**
   The main task of a tokenizer is to segment an input string into a sequence of token with sentential boundaries marked. We are using a Tokenizer developed by Helmut Schmid (IMS, Schmid (2000b)). The period disambiguation is solved to a satisfying extent, i.e. an accuracy of at least 99,5%. The determination of token boundaries is more problematic because many of the technical terms contain special signs (like brackets, colons etc) and may in addition be multiword expressions. To achieve a precision that also comes close to 100% a comprehensive lexicon of and a morphological component for biochemical terminology is needed. This is the task of the lexicon group GenIELex of GenIE. For details see Section 5.

   For the example above we can follow the simple heuristic that a blank marks a word boundary. Except in one case: *fructose 6-phosphate* has to be recognized as a multiword. This is done within the next processing step.

2. **Tagging and multiwords**
   The second processing step concerns part-of-speech (POS) tagging including lemmatization. POS-tagging is the process of going through a corpus of sentences and labelling each word in each sentence with its POS. Within GenIE we are using the TreeTagger developed by Helmut Schmid Schmid (1994). This tagger estimates transition probabilities using a decision tree. The implementation reached an accuracy of 96.36% on Penn-Treebank data. With respect to biochemistry related corpora the POS-tagging suffers the same lexical gap as the Tokenizing, and will be captured by GenIELex, too.

   The tagging process is followed by a correction step which eventually recognizes multiword terms. The relevant information for composition of these terms comes from several tools and resources, in particular a lexicon of multiword terms and an abbreviation detector (according to Roy (unknown)). The abbreviation *(Fru-6P)* is a significant trigger for recognizing *fructose 6-phosphate* as a multiword term.

3. **Chunking**
   Text chunking consists of dividing a text in syntactically correlated non-overlapping phrases (sentence parts). Text chunking is an intermediate step towards full parsing and is usually based on finite-state techniques. The considerable advantages a chunker has are that robustness and speed are primary design considerations. In addition precision is at a high level[3], too.
   We use Steven Abneys partial parser Cass[4]. It is an already approved system, that allows the user to modify the cascading set of rules and adjust the system to give it a seperate identity; which – in our case – is a biochemistry related identity.
   As long as incomplete results suffice, we will prefer chunking instead of full parsing[5].

   The chunking step leads to the following result:

   [$_{c}$ [$_{c0}$ [$_{np}$ Phosphofructokinase] [$_{vx}$ [$_{vbz}$ catalyzes]]]
   [$_{ng}$ [$_{nx}$ [$_{dt}$ the] [$_{nn}$ transfer]] [$_{of}$ of] [$_{nx}$ [$_{dt}$ the] [$_{nn}$ gamma-phosphate]]]
   [$_{pp}$ [$_{in}$ from] [$_{np}$ MgATP]]
   [$_{pp}$ [$_{to}$ to] [$_{nx}$ [$_{nn}$ fructose] [$_{nn}$ 6-phosphate]]]
   [$_{pp}$ [$_{in}$ in] [$_{ng}$ [$_{nx}$ [$_{dt}$ the] [$_{jj}$ first] [$_{jj}$ committed] [$_{nn}$ step]] [$_{of}$ of] [$_{nx}$ [$_{nn}$ glycolysis]]]]
   ]
   [*sent* .]

[3] Precision of more than 90% are the rule.
[4] www.research.att.comãbney
[5] This reflects our two-dimensional approach, where the leading idea is not to generate more structure than necessary.

4. **Full Parsing**

   LoPar is parser for probabilistic context-free grammars (PCFG) and head-lexicalised probabilistic context-free grammars (HPCFG) developed by Helmut Schmid[6] (IMS, University of Stuttgart). A PCFG is a context-free grammar which assigns a probability to each gammar rule. Such a grammar is able to rank different analysis of a sentence according to their probabilities. But, as PCFG fails to resolve syntactic ambiguities like PP-attachment, information about the lexical heads come into operation.
   The parser is fed with a grammar and a corpus and as an output it generates a syntactic language model which allows for parsing and disambiguation of the relevant sentences or phrases.

5. **Word sense (/Semantic) tagging**
   Within the step of semantic tagging, we annotate the chunks or partial parse trees with semantic tags. We basically distinguish between three types of semantic annotation. The first type annotates concepts of the underlying ontology (which serves as a backbone for semantic operations and as a logical interface between different components of the system.) The second type associates semantic tags to closed class words of English (such as, *of*, *not*, *and*, etc.) And the third type of semantic annotations is context dependent. Here the context may be given by a particular syntactic environment, eventually including semantic tags or coreferent phrases (and their semantics). (This type of annotation thus takes place at some point in the two dimensional structure of our system.) We show the results for this step by placing a semantic tag as subscript at the end of the bracket[7]:

   $[_c\ [_{c0}\ [_{np}$ Phosphofructokinase$]_{ENZ}\ [_{vx}\ [_{vbz}$ catalyzes$]_{ENZ-REACT}]]$
   $[_{ng}\ [_{nx}\ [_{dt}$ the$]\ [_{nn}$ transfer$]_{CH-REACT}]\ [_{of}$ of$]_{OF}\ [_{nx}\ [_{dt}$ the$]\ [_{nn}$ gamma-phosphate$]_{CH-ELEM}]]$
   $[_{pp}\ [_{in}$ from$]\ [_{np}$ MgATP$]_{CO}]$
   $[_{pp}\ [_{to}$ to$]\ [_{nx}\ [_{nn}$ fructose$]\ [_{nn}$ 6-phosphate$]]_{FRU}]$
   $[_{pp}\ [_{in}$ in$]\ [_{ng}\ [_{nx}\ [_{dt}$ the$]\ [_{jj}$ first$]\ [_{jj}$ committed$]\ [_{nn}$ step$]]\ [_{of}$ of$]_{OF}\ [_{nx}\ [_{nn}$ glycolysis$]_{PATHW}]]]$
   ]
   $[_{sent}$.]

   The tags are arranged in a way to form a taxonomy that reflects the underlying ontology (see below).

6. **Template detection with subcat information**
   To get the relevant information for the filling of the templates subcat frames have to be detected. The lexicon provides the following two necessary entries:

catalyze	$SUBJ_{nom}$: ENZ		OBJ_{acc}: CH-REACT	
transfer	PP_{of}:	CH-ELEM	PP_{from}:CO	PP_{to}:FRU

   The selectional restrictions are part of the lexical entries. Within this processing step the arguments have to retrieved and matched.

7. **Template filling**
   After matching the arguments the relevant information can be extracted with help of template filling rules, like the following:

A:	SUBJ:ENZ	$\Longrightarrow$	**Catalyze event** e_1
e_2:	OBJ:CHEM-REACT		Catalyst: A
e_1:	catalyze:ENZ-REACT		Reaction: e_2

[6] See Schmid (2000a)

[7] We are using the following abbreviatios as semantic tags: ENZ as abbreviation for Enzyme; CH-REACT classifies the kind of event as CHEMICAL REACTION; CH-ELEM as abbreviation for CHEMICAL ELEMENT; PATHW as abbreviation for PATHWAY; CO as abbreviation for ORGANIC COMPOUND and FRU abbreviates FRUCTOSE.

The extracted information can then be characterized by two events: the catalyze event e_1 and the transfer event e_2. The following table shows the filled template:

Catalyze event: e_1	**Transfer event**: e_2
Catalyst: Phosphofructokinase	Type: chem. reaction
Reaction: e_2	Transfer of: gamma-phosphate
	Transfer from: MgATP
	Transfer to: fructose 6-phosphate
	Transfer within: glycolysis

8. **Ontology**

The semantic tags, templates and the database system are all specified by the underlying ontology developed by EML.[8] An ontology is a description of concepts and relationships, that model a part of the world. What is important in this modelling process depends on what an ontology is for. One of our primary goals is the implementation of an ontology which models biochemical pathways and includes concepts ranging from basic biochemistry to genomics. So, as an example the ontology describes the fact, that anything belonging to the concept *DNA kind* has an *Intron* and an *Exon*[9] etc., and it is a member of the concept *Nucleic acid kind*. All properties that are ascribed to the concept *Nucleic acid kind* are also inherited to the concept *DNA kind*. Hence, as a nucleic acid consists of nucleotides anything belonging to *DNA kind* consists of nucleotides, too.

Altogether the ontology has the following duties:

- it establishes the backbone for inferences within the information extraction system (as illustrated already in Section 3),
- it represents joint terminology and logical interface between the lexicon, the database system and the extracted information, and
- it is relevant for the consistency of the extracted data and the database system.

The database system consists of (i) an Ontology Managemant System, i.e. a tool to build and manage ontologies based on a logic based language (KIF); (ii) a Deductive Database Generator, which generates a deductive database from the specification of the ontology and provides an Application Programming Interface (API); (iii) a deductive database which follows the semantics of the Ontology, is highly optimized, has a type system and allows for consistency checking.

3 Semantic tagging and deduction schemata

The following two sentences were taken from "The Journal of Biological Chemistry" (Byrnes et al. (1999)):

> *The serine protease CD26/dipeptidyl-peptidase IV (CD26/DPP IV) and chemokines are known key players in immunological processes. Surprisingly, CD26/DPP IV not only removed the expected Gly1-Pro2 dipeptide from the NH2 terminus of macrophage-derived chemokine (MDC) but subsequently also the Tyr3-Gly4 dipeptide, generating MDC(5-69).*

The first two columns of the following table represent the output of the first processing steps[10]: tokenizing and tagging. We will focus here on the third step, the semantic annotation (or semantic

[8] This work is being carried out together with Ontology Works, Hanover, Maryland U.S.A.
[9] Procaryotic DNA is regarded as a special case where Introns have legth 0.
[10] We only illustrate this for the second (relevant) sentence.

tagging)[11]. The first column shows one lemmatized token per line. The second column gives the corresponding POS tag and the third column shows the result of annotating semantic tags to tokens.[12] If no arrow points to a tag, this means that it has been annotated at this stage of processing by direct projection from the lexicon and without exploitation of any kind of contextual knowledge. The tags with arrows pointing to them at least depend on the knowledge at the source of the arrows. They indicate that some steps in the direction of deeper processing had to be achieved in order to identify the correct tag. In particular the fact that CD26/DPP IV is an Enzyme can only be annotated after its contextual resolution to the occurrence of CD26/DPP IV in the first sentence; and that MDC abbreviates 'macrophage-derived chemokine' depends on the semantic tag of chemokine and the particular syntactic environment indicated by the dependency on PP. Note that we did not semantically annotate MDC(5-69) at this stage, simply because it cannot be reliably identified as an abbreviation of a protein. (The similarity between MDC(5-69) and MDC depends on the meaning of (5-69). But as the meaning of phrases in parenthesis is highly ambiguous nothing may be predicted at this stage.) The results of chunking are given in column 5.

Lemma	POS	STAG	Chunks			STAG
Surprisingly,	RB	-				
CD26/DPP IV	NP	ENZ	NP			
not	RB	-				
only	RB	-				
remove	VBD	-		VP		ENZ-REACT
the	DT	-				
expect	JJ	-				
Gly1-Pro2	NN	PEP	NP			PEP
dipeptide	NN	PEP				
from	IN	-				
the	DT	-	NP			
NH2 terminus	NN	PEP-CH				
of	IN	OF		NP		PEP-CH
macrophage-derive	JJ	-	PP			
chemokine	NN	PROT				
(MDC)	NP	PRO-ABR				
but	CC	-				
subsequently	RB	-				
also	RB	-				
the	DT	-				
Tyr3-Gly4	NN	PEP	NP			PEP
dipeptide	NN	PEP				
generate	VBG	-	VP			
MDC(5-69)	NP	-	NP	VP		PRO-ABR

A next, intermediate step has as its task a kind of semantic preprocessing. This step deals mainly with semantic annotation of complex chunks. The semantic tag for the complete chunk has to be calculated from its parts, their semantic tags and the way they are composed. This is illustrated by the NP *the NH2 terminus of macrophage-derived chemokine*, to which a rule of

[11]Further abbreviatios as semantic tags: PEP as abbreviation for PEPTIDE; PEP-CH as abbreviation for PEPTIDE CHAIN; PRO as abbreviation for PROTEIN; PRO-ABBR as abbreviation for PROTEIN ABBREVIATION ; CC as abbreviation for CHEMICAL COMPOUND. The tags are arranged in a way to form a taxonomy that reflects the underlying ontology.

[12]The discussion of this section presupposes a unique, preferred output of the first processing steps. In principle we could give up this presupposition and reclaim the two-dimensional approach to include tagging as well as tokenizing. We didn't do this for two reasons. First, the results of the tokenizer and tagger are very good for standard English texts and will achieve this precision also for the biochemical texts once they are trained with the continuously increasing lexicon of biochemical terminlogy.

the form x:**PEP-CH OF** y:**PROT** → x:**PEP-CH** ∧ part_of(x,y) will be applied projecting the semantic tag of the head of the phrase and identifying the meaning of the preposition *of* as *part_of* relation. The meaning of the preposition depends on the semantic tags of its arguments, which both are **PEP-CH**. (For y this must be derived from the ontology: proteins are peptide chains.)

This kind of meaning dependency also has to be exploited in order to calculate the last two semantic tags for the verbs *remove* and *generate* in the table above. The most frequent verbs in our corpus (see Section 3) are verbs that belong to standard English and have in general a broad variety of senses or meanings. And most of them are different, or more general than the meanings they typically have in the biochemical domain. As these meanings may only be determined on the basis of the semantic types of arguments these verbs subcategorize for their semantic tag depends on the calculation of the (semantic tags in the) subcategorization frame. The relevant subcategorization frame of the verb *remove* is 'NP_nom **remove** NP_acc **from** NP_dat'. Its slots will be filled with the semantically tagged representations of the arguments, which we denote as CD26/DPP_IV:**ENZ**, Gly1-Pro2:**PEP**, and MDC:**PROT**. If we make explicit the referential argument of the verb, i.e. the event e_1 it describes, in a DRT style manner[13] we get the following condition.

(1) e_1: remove(CD26/DPP_IV:**ENZ**,Gly1-Pro2:**PEP**,MDC:**PROT**)

That *remove* describes a chemical, and in particular an enzymatic reaction can now be derived on the basis of the semantic types of its arguments. In general a chemical reaction involves a change in which the atoms or molecules of two or more substances, the *substrates* of the reaction, are rearranged to form one or more additional substances, the reaction *products*. If this change is brought about by an enzyme, or catalyzed by a protein we have an enzymatic reaction. In addition *remove* implies that a molecular part y is taken away from a substrate z. We use lexical rules like the following to distinguish between different senses of verbs on the basis of their argument types.

e:remove(x:**ENZ**,y:**CC**,z:**CC**) → ENZ-REACT(e) ∧ e: z := y ⊕ z\y

We use the notation ':=' to indicate that the chemical compound z has changed its composition to y ⊕ z\y. Here the operator ⊕ denotes the mereological sum operator[14], and the operator \ represents the result of cleavages.[15] Note that although we are using a variable e to denote the reaction event we do not explicitly specify the prestate and poststate of this event. The information that y is a molecular part of z during the prestate of e but no longer in the result state is implicitly given by the fact that z is a substrate of e but not a product. More precisely, a rule like

e: z := y ⊕ z\y → z ∈ Substrates(e) ∧ y ∈ Products(e)

will allow us to derive the following information that may be (part of/or) added to an enzyme database.

Enzyme name	Substrates	Products
CD26/DPP IV	MDC	Gly1-Pro2
		MDC\Gly1-Pro2

We now come to the cleavage of the Tyr3-Gly4 dipeptide which is expressed by an elliptical construction triggered by the phrase *not only ... but subsequently also*. We will not attempt do perform any syntax-based reconstruction of the elided phrase. We will instead use the semantic

[13] A thorough Introduction can be found in Kamp and Reyle (1993)

[14] The theory of mereology can be found in Varzi (1996).

[15] The authors of the article would represent MDC\Gly1-Pro2 instead as MDC(3-69), using parentheses the meaning of whose is highly ambiguous. We will come back to this shortly.

information that represents the source (and which is identified by the *not only* part of the trigger) in order to complete the semantics of the NP following *but subsequently also*.[16] Standard reconstruction of the elided phrase will extend the condition in (1) to the condition set in (2).

(2) e_1: remove(CD26/DPP_IV:**ENZ**,Gly1-Pro2:**PEP**,MDC:**PROT**)
e_1: remove(CD26/DPP_IV:**ENZ**,Try3-Gly4:**PEP**,MDC:**PROT**)

If we take into account the implications of the adverb *subsequently* we may consider $e_1 \circ\!\prec e_2$ as complex event consisting of the succession of e_1 and e_2. The properties of this complex event are determined by the general schema[17]

$$e_1 \circ\!\prec e_2 \wedge e_1\colon z := y_1 \oplus z \backslash y_1 \wedge e_2\colon z := y_2 \oplus z \backslash y_2$$
$$\rightarrow \quad e_1 \circ\!\prec e_2\colon z := y_2 \oplus (y_1 \oplus z \backslash y_1) \backslash y_2$$

which allows us to derive

Enzyme name	Substrates	Products
CD26/DPP IV	MDC	Gly1-Pro2 MDC\Gly1-Pro2
	MDC\Gly1-Pro2	Tyr3-Gly4 MDC\Gly1-Pro2-Tyr3-Gly4

From this the substrate specificity of CD26/DPP IV can be computed automatically to contain Gly1-Pro2 and Tyr3-Gly4.

The last event that will be added to (2) is e_3: generate(CD26/DPP IV, MDC(5-69)). For the verb *generate* we have the lexical rule

e:generate(x:**ENZ**,y:**CC**) →
ChemReact(e) ∧ agent(e):PROT ∧ y ∈ Products(e)

This event is understood as consisting of the sum of the two preceding events and thus producing the same result as they do, we may derive that MDC(5-69) must be equal to (MDC\Gly1-Pro2)\Tyr3-Gly4, or MDC\Gly1-Pro2-Tyr3-Gly4, and hence get an indication of what the parentheses in MDC(5-69) are supposed to mean.

4 Corpus Data and Lexicon Acquisition

Baayen (Baayen (2001)) reports the fact that in every corpus there are as much hapax legomena, i.e. words that occur only once in a corpus, as there are other word types. This shows that one cannot analyze arbitrary texts without a reliable word formation component.
Within GenIE the GenIELex group has its main task in developing a biochemistry specific lexicon as well as an annotated corpus for the evaluation. The basic necessity for building up such lexica is illustrated by the following figures reflecting some properties of our JBC corpus[18], too.

Tokens [19]	30 * 10^6
Unknown Tokens [20]	3.5 * 10^6
Types	405 000
Hapax Legomena (1)	205 000
Unknown Types (2)	322 000
(1) ∩ (2)	172 500

[16] There are several approaches to a semantic based reconstruction of ellipses. M. Dalrymple and Pereira (1991), and others use higher order unification; Asher (1993) exploits parallelism; Schiehlen (1999) (IMS, University of Stuttgart) has a very sophisticated algorithm

[17] The variables are interpreted dynamically, see Groenendijk and Stockhof (1991).

[18] We built up a full text corpus from articles of the *Journal of Biological Chemistry* - www.jbc.org. The corpus has about 30*10^6 token. This correspond to approximately 1/5 of the complete available Journal.

As the amount of hapax legomena in the corpus is 204.775 of which 172.458, i.e. 85%, are unknown they are very likely to belong to the biochemical terminology.

Taking into account an average length of a sentence in our JBC corpus of 27 words and the average of every 9th token being unknown it is almost sure to have more than one unknown token per sentence. Even worse, three unknown token per sentence are the average. Even more, the amount of hapax legomena (most of them are biochemistry related technical terms) show the basic need for at least semi-automatic methods of lexicon generation.

Lexicon

There seems to be a general agreement within the literature on term extraction that not only terms but also their contexts must be taken into consideration in order to provide tools for automatic extraction of terms and in particular for automatic abstracting information on the semantic role of terms. The notions of context taken into consideration are:

- lexical, e.g. *association between* $\langle NP \rangle$ *and* $\langle NP \rangle$
- syntactic, e.g. ADJ+NN, NN+*of*-NN, or V+NN collocations,
- paralinguistic, e.g. parentheses, and
- so-called "knowledge rich"[21] textual contexts, which indicate important conceptual characteristics for the search term, and are thus used for the acquisition of domain knowledge and for the construction of ontologies.

For the detection (and also for the extraction) of such frames we use the IMS Corpus Query Processor (CQP). CQP is one of the tools from the IMS Corpus Workbench[22]. It is a specialized search engine for linguistic research. Information can be extracted from text corpora by the use of the CQP query language and by some additional commands and parameters[23].

The acquisition process is recursive. We start with semantically tagged morphemes and compositionally calculate the semantics of expressions like *phosphorylation*, *carboxamidomethylation*; *dephosphorylation*, or *5'-(de)phosphorylation*, *Galactosylation*, or *glucosylation*. (See Gerstenberger (2001). Note that there are 576 one word term types (out of 43048 tokens) ending on *ylation* in our corpus.) This will then be used to enrich the lexicon and to enhance CQP work.

5 Summary and Outlook

We described a system (GenIE) that automatically extracts information about biochemical pathways from free text sources, i.e., research papers and comments found in existing databases.

The system combines sophisticated tools from computational linguistics. The major knowledge sources of the system are (i) a lexicon of biochemical terms and morphemes, (ii) a morphological analyzer for derivationally and compositionally complex expressions, (iii) a domain-specific ontology, comprising (iv) a theory of the atomic predicates and relations used in the lexicon. The items in the lexicon will be annotated with POS-tags, ontological tags, and their lexical semantic. The morphological analyzer will be able to compose each type of information associated with the parts to the corresponding information types associated with the compounds.

The construction of these resources is a still ongoing process which will be accomplished by developing and recursively applying algorithms for the semi-automatical aquisition of morphological, syntactical and collocational information to a corpus of biochemical texts (which at the moment comprises ca. 3×10^7 tokens). A core ontology has already be constructed and will be used in these algorithms, first, to be recursively extended and refined, and second, to fix the relationship between lexical semantics and ontological tagging.

[19] We applied a tagger/lemmatizer that had been trained on the Brown Corpus to the biochemical corpus.
[20] The token that weren't recognized by the lemmatizer.
[21] Meyer (2001)
[22] The IMS Corpus Workbench is a set of tools for the manipulation of large, linguistically annotated text corpora.
[23] CQP assumes that corpora have been indexed beforehand.

The second goal consists in realizing a dynamic two-dimensional approach of the extraction process. One dimension concerns the combination of shallow interpretation techniques and in-depth analyses of relevant information bits. The other dimension concerns the possibility to exploit domain specific ontological and semantical knowledge at any stage of morphological and syntactic processing, or vice versa.

The extraction process itself is mediated by a semantic representation of the text which is then used to fill the data base. This semantic representation will make contextual information explicit and will serve as a basis for disambiguation tasks and contextual resolution. The semantic representations (at any level) are required not to overspecify, i.e. not to force any ambiguity the resolution of which is irrelevant for the extraction task at hand.

References

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer Academic Publisher.

Baayen, H. (2001). *Word frequency distributionse*. Kluwer Academic Publishers. to appear.

Byrnes, W. M., Hu, W., Younathan, E. S., and Chang, S. H. (1999). A chimeric bacterial phosphofructokinase exhibits cooperativity in the absence of heterotropic regulation. *The Journal of Biological Chemistry*, 274:3988–3993.

Gerstenberger, C. (2001). Semantische analyse von namen organischer verbindungen. to appear.

Groenendijk, J. and Stockhof, M. (1991). Dynamic predicate logic. *Linguistics and Philosophy*, 14:39 – 100.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic*. Kluwer.

M. Dalrymple, S. M. S. and Pereira, F. C. N. (1991). Ellipsis and higherorder unification. *Linguistics and Philosophy*, 14:399–452.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In Didier Bourigault, Christian Jacquemin, M.-C. L., editor, *Recent advances Computational Terminology*, chapter 14, pages 279–302. Kluwer.

Proost, P., Struyf, S., Schols, D., Opdenakker, G., Sozzani, S., Allavena, P., Mantovani, A., Augustyns, K., Bal, G., Haemers, A., Scharpé, A.-M. L. S., Damme, J. V., and Meester, I. D. (1995). Truncation of macrophage-derived chemokine by cd26/ dipeptidyl-peptidase iv beyond its predicted cleavage site affects chemotactic activity and cc chemokine receptor 4 interaction. *The Journal of Biological Chemistry*, 24:3828–3835.

Roy, Y. P. Hybrid text mining for finding abbreviations and their definitions.

Schiehlen, M. (1999). *Semantikkonstruktion*. PhD thesis, Universität Stuttgart.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK. unknown.

Schmid, H. (2000a). Lopar: Design and implementation.

Schmid, H. (2000b). Unsupervised learning of period disambiguation for tokenisation.

Varzi, A. C. (1996). Parts, wholes, and part-whole relations: The prospects of mereotopology. *Data and Knowledge Engineering*, 20:259–286.

Protein Functional Classification by Text Data-Mining

B.J. Stapley *, L.A. Kelley and M.J.E. Sternberg †
Biomolecular Modelling Laboratory
Imperial Cancer Research Fund
London, UK
(b.stapley|l.kelley|m.sternberg)@icrf.icnet.uk

Abstract

Classifying proteins into classes based on their cellular rôle or function is a powerful way of making sense of genomic data. Here we present a method for the automatic classification of proteins by analysis of relevant medline abstracts. We employ keyword matching using a thesaurus of *S.cerevisiae* gene naming terms to retrieve relevant text for each protein. From these term vectors are generated for each protein. We then train support vector machines to automatically partition the term space and to thus discriminate the textual features that define a protein's function. We test the method on the task of assigning proteins to various subcellular locations. The method is benchmarked on a set of proteins of known sub-cellular location. No prior knowledge of the problem domain nor any natural language processing is used at any stage. Our method has comparable performance to rule-based text classifiers and we find that amino acid compositional information is a poor predictor when employed in isolation. Combining text with protein amino-acid composition improves recall. We discuss the generality of the method and its potential application to a variety of biological classification problems.

Keywords: Support vector machine; Text classification; Subcellular localization; Proteins.

1 Introduction

Classifications are extrinsically valuable and have been at the heart of biological science since the 18^{th} century (Linnaeus, 1748). A classification can help organise and rationalise information, allowing the biologist to rapidly retrieve, compare and visualise entities and groups of entities. A classification also permits the rapid assimilation of new information into an existing framework. Examples of biological classifications include the Enzyme Commission Registry (Bairoch, 1994), and the MIPS classification (Mewes et al., 1999).

More recently, the concept of a classification has been extended to that of an ontology - a complete conceptualisation of field of study (Karp, 2000). Such ontologies include the EcoCyte project (Karp et al., 2000) and Gene Ontology project (Consortium., 2000). The aim of such projects is not just to systematically organise biological information for user querying but also to permit automatic knowledge discovery (Karp, 2001).

One important aspect of protein function that can be readily classified is the location of that protein within the cell. In order to carry out its physiological role, a protein must often be proximal to other components involved in that process; thus knowledge of sub-cellular localization can restrict the number of possible processes with which a protein can be involved. Location can also alter the experimental approach to characterizing a protein - e.g. purification.

*present address; Biomolecular Sciences, University of Manchester Institute of Science & Technology, PO Box 88, Manchester, UK, M60 1QD

†present address; Department of Biological Sciences, Imperial College of Science, Technology and Medicine, London, SW7 2AY, United Kingdom

It has proved surprisingly difficult to automatically predict protein cellular location from sequence alone (Eisenhaber and Bork, 1998). It is been know that the amino acid composition of protein can be an indicator of its sub-cellular location (Nishikawa and Ooi, 1982). It is also clear that many cellular compartments have proteins assigned to them according to targeting signals within the protein sequences; however, such signals are not universal or necessarily clearly defined.

For proteins which have been partially annotated/characterized, an alternative approach, pioneered by Eisenhaber and Bork is to use the existing textual information relevant to a protein to classify it to a particular sub-cellular location (Eisenhaber and Bork, 1999). They have developed a method called Meta-Annotator that classifies annotated proteins in SWISS-PROT using a set of manually generated biological rules (Bairoch and Apweiler, 1999). After tokenizing the annotations the rules are applied and a sub-cellular location extracted/predicted. The method is a great improvement over simple matching of relevant keywords within the documents. The authors report that 88% of SWISS-PROT entries can been assigned to a cellular compartment by this method as opposed to the 22% that be achieved by key-word matching.

Despite the success of Eisenhaber and Bork's technique, it has two inherent weaknesses: first, a set of rules must be generated - this is obviously less intensive than manually classifying the documents, but is subjective and costly in time; second, in order to tokenize the documents they must already be structured - free text cannot be treated in such a manner. The method described in this paper is to treat the protein as a vector of terms from relevant Medline documents. This approach derives from the vector-based model common in information retrieval (van Rijsbergen, 1979). The term weights of a vector are a function of their frequencies within the document collection as a whole and the frequency within the relevant documents. Given a set of protein term-vectors the task is to find some function that partitions the space according to the localisation of the protein. For this task we employ support vector machines (SVM) (Vapnik, 1995).

Support vector machines (Vapnik, 1995) are a computational method for performing simultaneous feature space reduction and binary classification based on Vapnik's statistical learning theory. SVMs have been applied to the problems of pattern recognition (Burges, 1998), regression estimation (Vapnik et al., 1997), and text-categorization (Cooley, 1999; Joachims, 1998; Kwok, 1999). Because SVMs cope well with high dimensionality and are very fast to train, they are particularly suited to problems in text data-mining/information retrieval. Kwok studied the use of SVMs in text categorization of Reuters newswire documents (Kwok, 1999). Cooley and Joachims have also carried out similar studies (Cooley, 1999; Joachims, 1998). Joachims also demonstrated that transductive inference can enhance classification when the training set is small (Joachims, 1999). Kwok suggests that performing single value decomposition/latent semantic indexing on the resulting support vectors and re-learning can improve performance (Kwok, 1999). Kwok, Cooley and Joachims have all demonstrated that SVMs can outperform other techniques in classifying text documents.

We evaluate the performance of SVMs in classifying a set of proteins of known sub-cellular locations from *S. cerevisiae*. Text relevant to these proteins is obtained from Medline by key-word matching of the gene naming terms. SVMs trained on the resulting term vectors classify the proteins with good precision and recall. We also show that SVMs trained on amino acid compositions are out-performed by our SVMs trained to text data and that combining amino acid composition and term vectors can enhance classification for some sub-cellular locations.

2 Methods

2.1 Document and term processing

Term vector representations of *cerevisiae* protein's were generated as follows. First, we scanned 22517 Medline documents for occurrences of yeast gene naming terms using a thesaurus. This thesaurus was obtained from the Saccharomyces Genome Database gene registry (Cherry et al., 2000) [1]. For each protein, any document that contained an occurrence of the gene name or aliases of that gene was considered relevant. We then remove very common words and reduced

[1] http://genome-www.stanford.edu/Saccharomyces/registry.html

the remaining words to just their stems (Porter, 1980). The term representation of a gene is a function of the number of relevant Medline documents and the occurrence statistics of the terms. We employed a variant of inverse document frequency (IDF) that takes account of the number of Medline documents relevant to a particular gene. The weight of term i for gene k is given by :

$$\log(1 + \sum_j f_j(w_i)) - \log N(w_i) - \log(1 + R_k) \quad (1)$$

where $f_j(w_i)$ is the frequency of term i in document j, $N(w_i)$ is the number of documents containing term i, and R_k is the number of medline documents relevant to gene k. Cooley suggests that the specific nature of term weighting may not not be crucial to the performance of SVMs in text classification (Cooley, 1999)

2.2 Classification

The assignment of yeast proteins to sub-cellular compartments was obtained from the MIPS web site [2]. According to MIPS, 2233 *cerevisiae* proteins have known locations in one or more of 16 categories. We limit our test and training data to these proteins. The locations and numbers of proteins at each location is shown in 1. For each location class, our training set consisted of half the number of genes that fall into this category plus half the of the remaining negative examples. The test set consists of the remaining proteins - positive and negative cases.

Table 1: Number of positive examples in training and test sets for sub-cellular location

Role/location	+ve in training set	+ve in test set
organisation of plasma membrane	67	63
organisation of cytoplasm	279	245
organisation of cytoskeleton	47	52
organisation of endoplasmatic reticulum	68	80
organisation of Golgi	44	33
nuclear organisation	267	341
organisation of chromosome structure	19	18
mitochondrial organisation	174	155
peroxisomal organisation	19	12
vacuolar and lysosomal organisation	27	16
extracellular/secretion proteins	10	5

2.3 Training of SVM's

We used the support vector machine program SVM Light package v3.50 (Joachims, 1998) [3]. We trained a SVM for each classification using a linear kernel function with the regulization parameter C calculated as $\frac{1}{mean(x \cdot x)}$.

2.4 Evaluation

We evaluate the classification performance using a variety of methods. For traditional text retrieval, evaluation measures based on precision/recall have been widely used (Baeza-Yates and Ribeiro-Neto, 1999). Precision is defined as $a/(a+b)$ and recall as $a/(a+c)$, where a, b, and c are the number of true positives, the number of false positives, and the number of false negatives, respectively. We use precision/recall plots calculated on the distance of each test vector

[2] available from http://mips.gsf.de/proj/yeast/catalogues/subcell/index.html
[3] available from http://ais.gmd.de/thorsten/svm_light

from the SVM decision boundary; however, comparison of performance between them is difficult because the classes contain different numbers of positive examples. To assess the global performance of classification methods we employed micro- and macro- averaging of the precision/recall data. Micro-averaging determines precision and recall of a set of binary classifiers averaged over the number of documents; this equates to evaluating average performance a document selected randomly from the test collection. In macro-averaging, the recall/precision are averaged over the number of classes. Macro-averaging estimates the expected performance of an SVM trained on a new class; whereas micro-averaging estimates the performance of the system with new documents. For our purposes, micro-averaging is more useful.

We also use the F1 measure proposed by van Rijsbergen (van Rijsbergen, 1979). F1 is given by $\frac{2rp}{r+p}$ where p and r are precision and recall respectively. We determine the maximal value of F1 for the performance of each system on a particular classification.

3 Results

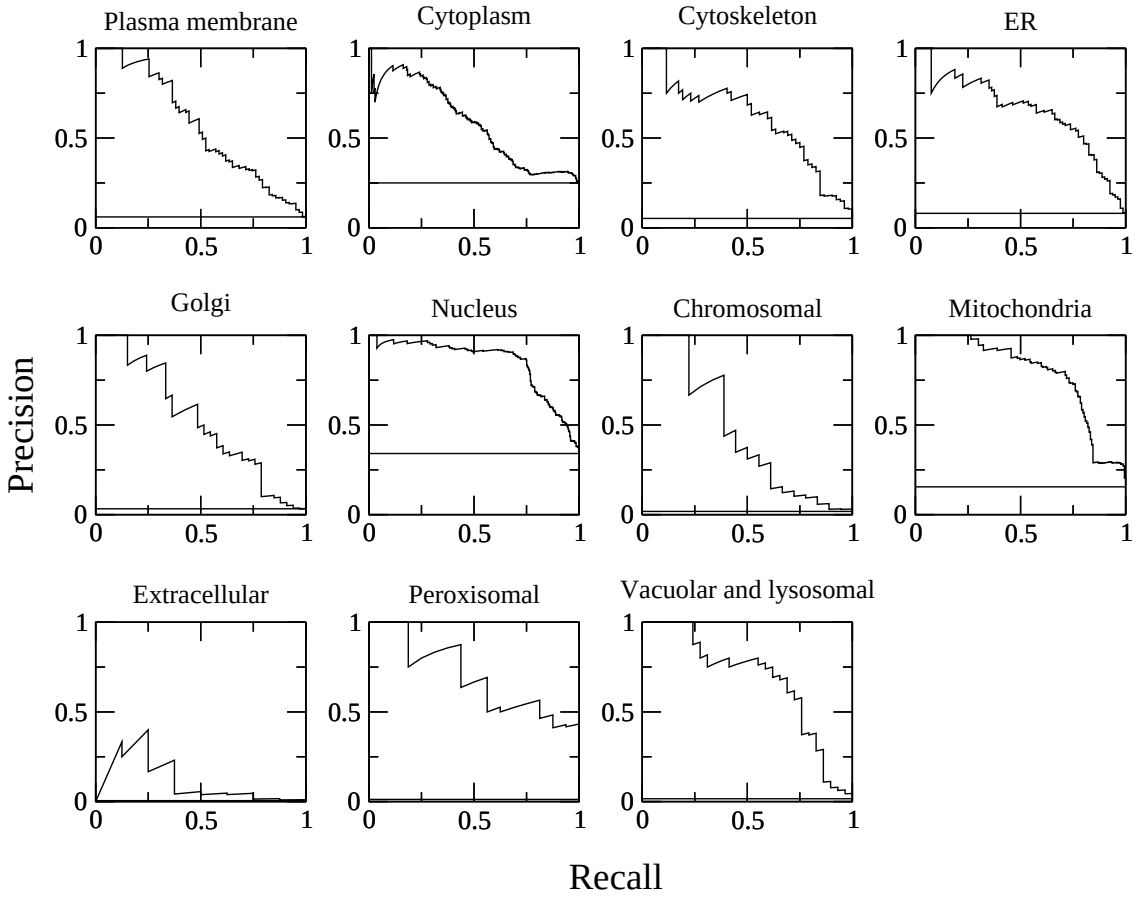


Figure 1: Precision/recall plots for location classifiers trained on term vectors. Horizontal lines indicate the performance of a random classifier

Precision/recall graphs for the various classifications are shown in figure 1. The performance of a random classifier is shown as a horizontal line in each plot. At low levels of recall, the precision is generally very high (95%+). Classes with a large number of positive examples - nuclear, cytoplasmic, and mitochondrial - are better predicted than the rarer classifications. This is reflected in averaged precision/recall shown in figure 2. The better apparent performance from

micro-averaging is a result of better prediction of bigger classes. The values of $Max(F1)$ are shown

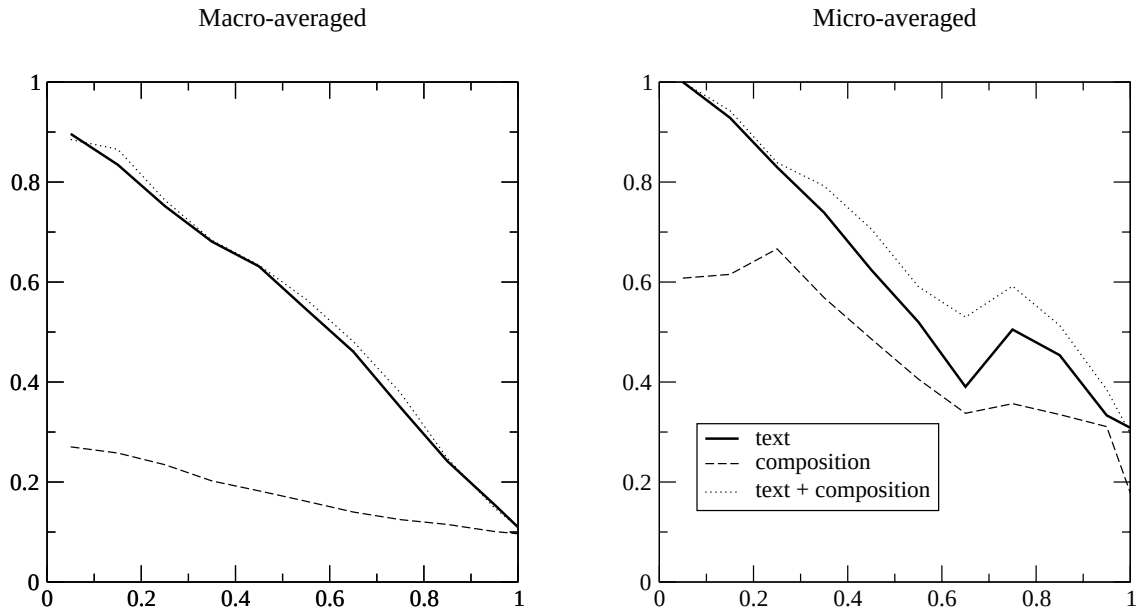


Figure 2: Micro and macro averaging of classification to 11 locational categories.

3.1 Sequence and text together improve classification

It has been known for some time that the amino acid composition of a protein can be used as an indicator of its sub-cellular localisation (Cedano et al., 1997; Nishikawa and Ooi, 1982). In particular, nuclear proteins generally contain disproportionally more charged and polar residues. Membrane associated proteins tend towards hydrophobicity, while intra-cellular proteins tend to be low in cysteine and rich in aliphatic and charged amino acids.

Figure 3 shows the performance of support vector machines in discriminating protein localisation based of their fractional composition of the twenty amino acids. It can been seen that composition is a poor predictor of ER, cytoskeleton, golgi, peroxisomal and vacuolar proteins, but good at predicting cytoplasmic, membrane and nuclear proteins. Composition also contains limited information on mitochondrial and chromosomal proteins. For extra-cellular proteins the scarcity of data makes assessment difficult, but the composition of these proteins gives better than random predictions. These results are in broad agreement with the recent work of Hua and Sun Hua and Sun (2001); a comparison with their work is included in the Discussion section of this paper.

Combining features from text with those of amino acid composition improves performance in classifying proteins to the cytosol and nucleus (table 2). In particular, recall is improved. This may reflect improved performance on those proteins which have relatively few citations in the literature.

3.2 Detecting errors in annotation

Any manual method of gene annotation is liable to errors of omission and mis-classification. We checked apparent false negatives and positives to assess whether they were genuine by inspection of the relevant Medline documents. We uncovered many examples of incorrect protein classification within MIPS, including several where proteins exist in both soluble and membrane associated forms. We also found numerous proteins which MIPS assigns to the nucleus that our method

Table 2: Maximum F1-value for classifications

Role/location	Max F1			
	text alone	text + composition	composition alone	random
organisation of plasma membrane	0.54	0.56	0.47	0.12
organisation of cytoplasm	0.55	0.60	0.48	0.39
organisation of cytoskeleton	0.62	0.61	0.13	0.10
organisation of endoplasmatic reticulum	0.65	0.66	0.10	0.14
organisation of Golgi	0.54	0.53	0.10	0.14
nuclear organisation	0.80	0.82	0.61	0.51
organisation of chromosome structure	0.52	0.52	0.20	0.04
mitochondrial organisation	0.75	0.75	0.36	0.27
peroxisomal organisation	0.67	0.65	0.03	0.01
vacuolar and lysosomal organisation	0.69	0.69	0.06	0.02
extracellular/secretion proteins	0.31	0.33	0.12	0.01

correctly flags as being non-nuclear. These include: UBC6 - a ubiquitin-conjugating enzyme, anchored in the ER membrane with the catalytically active domain in cytoplasm (Lenk and Sommer, 2000); hts1 - a histidyl-tRNA synthetase which is located exclusively in the mitochondria and cytosol (Chiu et al., 1992); and SMI1 protein involved in beta-1,3-glucan synthesis which has been shown to localise in patches at bud sites (Martin et al., 1999). Thus automatic functional assignment of proteins can be used to improve manual assignment by spotting errors and increasing recall.

For the cytosolic classification, the top scoring 'false' positive is cdc42, a Rho-type GTPase involved in bud site assembly and cell polarity. cdc42 contains a CAAX motif for geranylgeranyl modification and is likely to be associated with cell membranes. Ziman et al. (1993) determined that cdc42 exists in both a soluble form and membrane associated form within the cell; thus cdc42 should be included in cytosolic classification. A similar situation exists with ypt1 which is a GTP-binding protein required for vesicle transport from ER to Golgi and within the Golgi stack. It also undergoes geranylgeranyl modification, but the abundance and significance of any cytosolic form of the protein is not clear.

4 Discussion

4.1 Functional classification using SVMs and text

The utitility of support vector machines in text classification tasks is now well established and in this paper we have demonstrated that the technique can be applied to the functional classification of proteins. Other than a list of gene naming terms and synonyms, our method uses no prior knowledge of the problem domain nor any information from previously compiled sequence databases. Much of the power of the method arises from 'detection' of co-cited documents which is probably a strong indicator of common function (Data not shown). We also found that use of kernel functions other than a first order polynominial offered very little significant increase in classification performance.

Automatic functional assignment of proteins can be used to improve manual assignment by spotting errors. Such errors may be simple mistakes, or the result of partial or incorrect information or understanding on the part of the human classifier. Even in the absence of such errors, assessments of what constitutes a correct assignment of documents into a classification will vary from user to user; thus there is a theoretical limit to the precision of an automatic classifier. What this limit may be is unclear but for nuclear and mitochondrial proteins, automatic classification

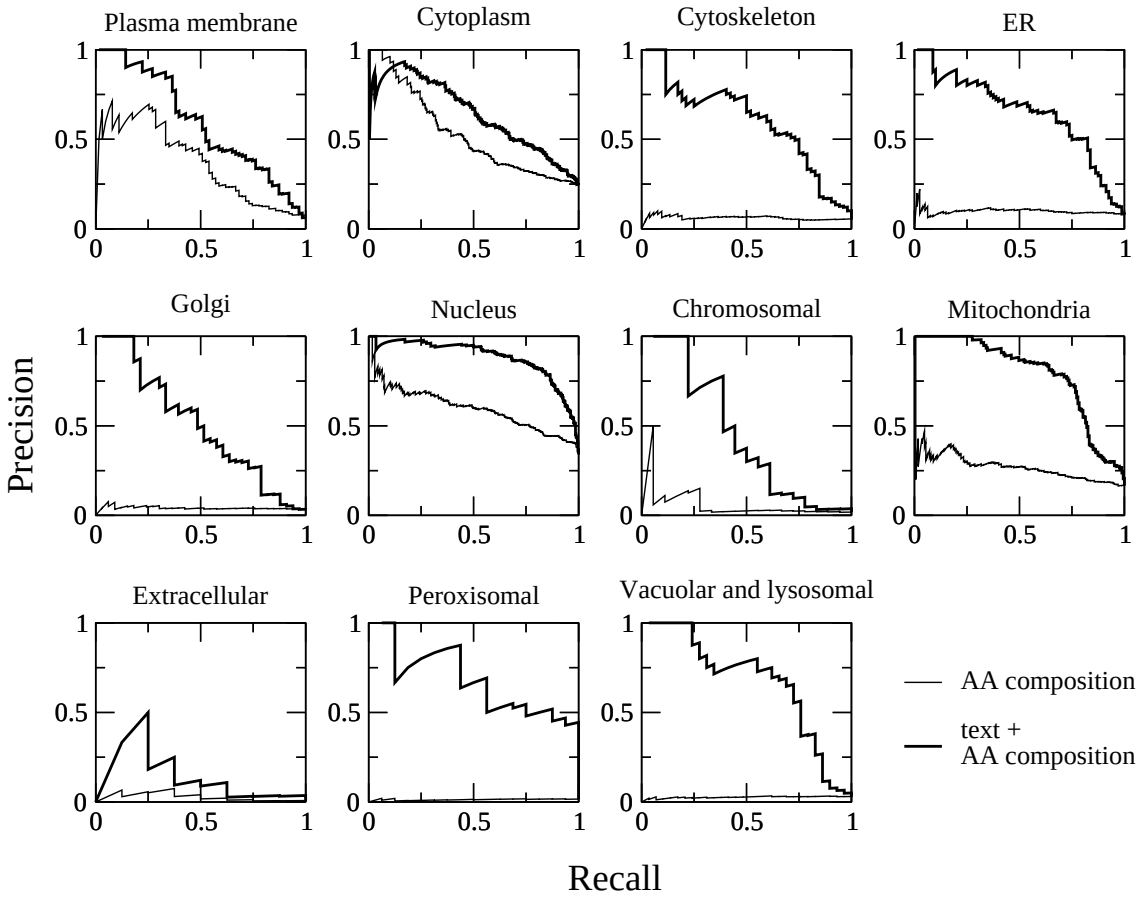


Figure 3: Precision/recall plots for location classifiers trained on term and/or amino acid composition vectors.

may be approaching this limit. This phenomenon indicates that discrete classifications may be an over-simplification and that they may limit the utility of classifications and ontologies for automatic knowledge discovery. Methods that incorporate uncertainty or fuzziness may be more applicable to such problems - although they are computationally more complex.

While this paper was in preparation a study of the use of support vector machines in predicting protein subcellular localization has appeared. Hua and Sun used amino acid compositions as features to train SVMs to discriminating between eukaryote cytoplasmic, extracellular, mitochondrial and nuclear proteins. Their test set consisted of 2472 eukaryote SWISS-PROT sequences sharing less than 90% pairwise sequence identity. The micro-averaged prediction accuracy was 79% using a radial basis function kernel and calculated using a jackknife test. A direct comparison of the two studies is not possible because of differences in kernel function, regularization parameter (C) and test set size and composition. The order of classification performance is conserved with nuclear and cytosolic proteins being easier to predict than mitochondrial.

For cellular compartments other than nuclear, cytoplasmic, or extracellular amino acid composition is generally a poor indicator. Unfortunately, relatively little text information is available in these rarer cases, so analysis of text fails to improve matters. However, where sufficient text information is retrievable, combining text and composition features can enhance performance.

4.2 Comparison with other text based methods

To compare our classification methods to that of Eisenhaber and Bork, we tested their algorithm (Meta-Annotator) on a subset of our original data that is present in SWISS-PROT (Bairoch and Apweiler, 1999). It should be borne in mind when comparing the two approaches that Meta-Annotator involves a large amount of manual intervention. Not only is the method only applicable to a previously manually curated protein database (Swiss-Prot), but it also has encoded into more than 1000 logical rules derived from a human expert. Our approach requires no human input other than a list of gene names and synonyms. Given these facts, it is little wonder that Meta-Annotator can generally out-perform our method. It is encouraging that a generic automatic approach can perform so well. With a larger set of training documents the SVM approach may be improved.

Meta-Annotator is outstandingly good at predicting mitochondrial proteins and very good at predicting nuclear proteins but less so with other locations. Because Meta-Annotator joins the golgi and endoplasmic reticulum (ER) into a single class, we modified our treatment of this locational class. A single SVM trained to distinguish golgi or ER from others performed very poorly, probably because the intersection of these two sets is very small (8 cases) according to the MIPS classification. We therefore used the max(F1) value from micro-averaging of two SVMs trained on the ER and golgi proteins independently. Text classification using SVMs out-performs Meta-Annotator for cytoplasmic and golgi/ER proteins.

Table 3: Comparison of text SVMs and Meta-Annotator

Role/location	Meta-A precision/recall	Meta-A F1	max(F1) for text
organisation of cytoplasm	49/32	0.38	0.54
organisation of Golgi/ER	75/48	0.58	0.62
nuclear organisation	87/86	0.86	0.80
mitochondrial organisation	90/93	0.91	0.75

4.3 Combining features for functional classification of proteins

Combining disparate features of a protein can aid in the functional classification of that protein - as demonstrated by Drawid and Gerstein (2000). With the advent of many high-throughput studies of genes and proteins, many more features can be used as training data for binary classifiers. These include protein interaction data, 2-D gel data, features of the protein or DNA sequence, and RNA expression array data. The inclusion of a variety of independent or semi-independent features should improve recall since data for every protein may not be available from every experiment. For example, our method can be applied to proteins/genes of unknown sequence or conversely, sequence information can be used to infer function in the absence of any text relevant to the protein/gene.

SVMs are highly suited to classification tasks of high dimensionality with many noisy or irrelevant features. There is little doubt that data from expression array and protein interaction experiments can yield insights into gene function, but the quality of such data is hard to determine. The method presented here may ameliorate some of these problems by automatically combining functional information present in the biomedical literature with experimental data.

References

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern information retrieval*. Addison-Wesley, Harlow, England.

Bairoch, A. (1994). The enzyme databank. *Nucleic Acids Research*, 22:3626–3627.

Bairoch, A. and Apweiler, R. (1999). The protein sequence data bank and its supplement trembl in 1999. *Nucleic Acids Res*, 27(1):49–54.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Cedano, J., Aloy, P., Perez-Pons, J. A., and Querol, E. (1997). Relation between amino acid composition and cellular location of proteins. *J Mol Biol*, 266(3):594–600.

Cherry, J. M., Ball, C., Dolinski, K., Dwight, S., Harris, M., Matese, J. C., Sherlock, G., Jin, G. B. H., Weng, S., and Botstein, D. (2000). Saccharomyces genome database. ftp://genome-ftp.stanford.edu/pub/yeast/SacchDB/.

Chiu, M. I., Mason, T. L., and Fink, G. R. (1992). Hts1 encodes both the cytoplasmic and mitochondrial histidyl-trna synthetase of saccharomyces cerevisiae: mutations alter the specificity of compartmentation. *Genetics*, 132(4):987–1001.

Consortium., T. G. O. (2000). Gene ontolgy: tool for the unification of biology. *Nature Genetics*, 25:25–29.

Cooley, R. (1999). Classication of news stories using support vector machines. In *International Joint Conference on Articial Intelligence Text Mining Workshop*.

Drawid, A. and Gerstein, M. (2000). A bayesian system integrating expression data with sequence patterns for localizing protein: comprehensive application to the yeast genome. *J. Mol. Biol.*, 301:1059–1075.

Eisenhaber, F. and Bork, P. (1998). Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol*, 8(4):169–70.

Eisenhaber, F. and Bork, P. (1999). Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics*, 15(7-8):528–35.

Hua, S. and Sun, Z. (2001). Support vector machine approach for protein subcellular locatization prediction. *Bioinformatics*, 17:721–728.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *10th European Conference on Machine Learning*, pages 137–142, Chemnitz, DE. Springer Verlag, Heidelberg, DE.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proc. 16th International Conf. on Machine Learning*, pages 200–209. Morgan Kaufmann, San Francisco, CA.

Karp, P. D. (2000). An ontology for biological function based on molecular interactions. *Bioinformatics*, 16(3):269–85.

Karp, P. D. (2001). Pathway databases: a case study in computational symbolic theories. *Science*, 293(5537):2040–4.

Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M., and Pellegrini-Toole, A. (2000). The ecocyc and metacyc databases. *Nucleic Acids Res*, 28(1):56–9.

Kwok, J. T.-Y. (1999). Automated text categorization using support vector machine. In *Proceedings of the International Conference on Neural Information Processing (ICONIP)*, pages 347–351, Kitakyushu, Japan.

Lenk, U. and Sommer, T. (2000). Ubiquitin-mediated proteolysis of a short-lived regulatory protein depends on its cellular localization. *J Biol Chem*, 275(50):39403–10.

Linnaeus, C. (1748). *Systema naturae, sive regna tria naturae systematice proposita per classes, ordines, genera, species ...* Stockholm, 6th edition.

Martin, H., Dagkessamanskaia, A., Satchanska, G., Dallies, N., and Francois, J. (1999). Knr4, a suppressor of saccharomyces cerevisiae cwh mutants, is involved in the transcriptional control of chitin synthase genes. *Microbiology*, 145:249–58.

Mewes, H. W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., and Frishman, D. (1999). Mips: a database for genomes and protein sequences. *Nucleic Acids Res*, 27(1):44–8.

Nishikawa, K. and Ooi, T. (1982). Correlation of the amino acid composition of a protein to its structural and biological characters. *J Biochem (Tokyo)*, 91(5):1821–4.

Porter, M. (1980). An algorithm for suffix stemming. *Program*, 14:130–137.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths, 2nd edition.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer Verlag, Heidelberg, DE.

Vapnik, V., Golowich, S., and Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. In Mozer, M., Jordan, M., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9, pages 281–287.

Ziman, M., Preuss, D., O., J. M., 'Brien, J. M., Botstein, D., and Johnson, D. I. (1993). Subcellular localization of cdc42p, a saccharomyces cerevisiae-binding protein involved in the control of cell polarity. *Mol Biol Cell*, 4(12):1307–16.

Twente Workshops on Language Technology

The TWLT workshops are organised by the PARLEVINK project of the University of Twente. The first workshop was held in Enschede, the Netherlands on March 22, 1991. The workshop was attended by about 40 participants. The contents of the proceedings are given below.

Proceedings Twente Workshop on Language Technology 1 (TWLT 1)
Tomita's Algorithm: Extensions and Applications
Eds. R. Heemels, A. Nijholt & K. Sikkel, 103 pages.

Preface and Contents

A. Nijholt *(University of Twente, Enschede.)* (Generalised) LR Parsing: From Knuth to Tomita.

R. Leermakers *(Philips Research Labs, Eindhoven.)* Recursive Ascent Parsing.

H. Harkema & M. Tomita *(University of Twente, Enschede & Carnegie Mellon University, Pittsburgh)* A Parsing Algorithm for Non-Deterministic Context-Sensitive Languages.

G.J. van der Steen *(Vleermuis Software Research, Utrecht)* Unrestricted On-Line Parsing and Transduction with Graph Structured Stacks.

J. Rekers & W. Koorn *(CWI, Amsterdam & University of Amsterdam, Amsterdam)* Substring Parsing for Arbitrary Context-Free Grammars.

T. Vosse *(NICI, Nijmegen.)* Detection and Correction of Morpho-Syntactic Errors in Shift-Reduce Parsing.

R. Heemels *(Océ Nederland, Venlo)* Tomita's Algorithm in Practical Applications.

M. Lankhorst *(University of Twente, Enschede)* An Empirical Comparison of Generalised LR Tables.

K. Sikkel *(University of Twente, Enschede)* Bottom-Up Parallellization of Tomita's Algorithm.

The second workshop in the series (TWLT 2) has been held on November 20, 1991. The workshop was attended by more than 70 researchers from industry and university. The contents of the proceedings are given below.

Proceedings Twente Workshop on Language Technology 2 (TWLT 2)
Linguistic Engineering: Tools and Products.
Eds. H.J. op den Akker, A. Nijholt & W. ter Stal, 115 pages.

Preface and Contents

A. Nijholt *(University of Twente, Enschede)* Linguistic Engineering: A Survey.

B. van Bakel *(University of Nijmegen, Nijmegen)* Semantic Analysis of Chemical Texts.

G.J. van der Steen & A.J. Dijenborgh *(Vleermuis Software Research, Utrecht)* Lingware: The Translation Tools of the Future.

T. Vosse *(NICI, Nijmegen)* Detecting and Correcting Morpho-syntactic Errors in Real Texts.

C. Barkey *(TNO/ITI, Delft)* Indexing Large Quantities of Documents Using Computational Linguistics.

A. van Rijn *(CIAD/Delft University of Technology, Delft)* A Natural Language Interface for a Flexible Assembly Cell.

J. Honig *(Delft University of Technology, Delft)* Using Deltra in Natural Language Front-ends.
J. Odijk *(Philips Research Labs, Eindhoven)* The Automatic Translation System ROSETTA3.
D. van den Akker *(IBM Research, Amsterdam)* Language Technology at IBM Nederland.
M.-J. Nederhof, C.H.A. Koster, C. Dekkers & A. van Zwol *(University of Nijmegen, Nijmegen)* The Grammar Workbench: A First Step Toward Lingware Engineering.

The third workshop in the series (TWLT 3) was held on May 12 and 13, 1992. Contrary to the previous workshops it had an international character with eighty participants from the U.S.A., India, Great Britain, Ireland, Italy, Germany, France, Belgium and the Netherlands. The proceedings were available at the workshop. The contents of the proceedings are given below.

Proceedings Twente Workshop on Language Technology 3 (TWLT 3)
Connectionism and Natural Language Processing.
Eds. M.F.J. Drossaers & A. Nijholt, 142 pages.

Preface and Contents
L.P.J. Veelenturf *(University of Twente, Enschede)* Representation of Spoken Words in a Self-Organising Neural Net.
P. Wittenburg & U. H. Frauenfelder *(Max-Planck Institute, Nijmegen)* Modelling the Human Mental Lexicon with Self-Organising Feature Maps.
A.J.M.M. Weijters & J. Thole *(University of Limburg, Maastricht)* Speech Synthesis with Artificial Neural Networks.
W. Daelemans & A. van den Bosch *(Tilburg University, Tilburg)* Generalisation Performance of Back Propagation Learning on a Syllabification Task.
E.-J. van der Linden & W. Kraaij *(Tilburg University, Tilburg)* Representation of Idioms in Connectionist Models.
J.C. Scholtes *(University of Amsterdam, Amsterdam)* Neural Data Oriented Parsing.
E.F. Tjong Kim Sang *(University of Groningen, Groningen)* A connectionist Representation for Phrase Structures.
M.F.J. Drossaers *(University of Twente, Enschede)* Hopfield Models as Neural-Network Acceptors.
P. Wyard *(British Telecom, Ipswich)* A Single Layer Higher Order Neural Net and its Application to Grammar Recognition.
N.E. Sharkey & A.J.C. Sharkey *(University of Exeter, Exeter)* A Modular Design for Connectionist Parsing.
R. Reilly *(University College, Dublin)* An Exploration of Clause Boundary Effects in SRN Representations.
S.M. Lucas *(University of Essex, Colchester)* Syntactic Neural Networks for Natural Language Processing.
R. Miikkulainen *(University of Texas, Austin)* DISCERN: A Distributed Neural Network Model of Script Processing and Memory.

The fourth workshop in the series has been held on September 23, 1992. The theme of this workshop was ”Pragmatics in Language Technology”. Its aim was to bring together the several approaches to this subject: philosophical, linguistic and logic. The workshop was visited by more

than 50 researchers in these fields, together with several computer scientists. The contents of the proceedings are given below.

Proceedings Twente Workshop on Language Technology 4 (TWLT 4)

Pragmatics in Language Technology

Eds. D. Nauta, A. Nijholt & J. Schaake, 114 pages.

The fifth workshop in the series took place on 3 and 4 June 1993. It was devoted to the topic "Natural Language Interfaces". The aim was to provide an international platform for commerce, technology and science to present the advances and current state of the art in this area of research.

Proceedings Twente Workshop on Language Technology 5 (TWLT 5)

Natural Language Interfaces

Eds. F.M.G. de Jong & A. Nijholt, 124 pages.

D. Speelman *(University of Leuven)* A Natural Language Interface that Uses Generalised Quantifiers.
R.-J. Beun *(IPO, Eindhoven)* The DENK Program: Modeling Pragmatics in Natural Language Interfaces.
W. Menzel *(University of Hamburg)* ASL: Architectures for Speech and Language Processing
C. Huls & E. Bos *(NICI, Nijmegen)* EDWARD: A Multimodal Interface.
G. Neumann *(University of Saarbrcken)* Design Principles of the DISCO system.
O. Stock & C. Strapparava *(IRST, Trento)* NL-Based Interaction in a Multimodal Environment.

The sixth workshop in the series took place on 16 and 17 December 1993. It was devoted to the topic ”Natural Language Parsing”. The aim was to provide an international platform for technology and science to present the advances and current state of the art in this area of research, in particular research that aims at analysing real-world text and real-world speech and keyboard input.

Proceedings Twente Workshop on Language Technology 6 (TWLT 6)
Natural Language Parsing: Methods and Formalisms
Eds. K. Sikkel & A. Nijholt, 190 pages.

Preface and Contents
A. Nijholt *(University of Twente)* Natural Language Parsing: An Introduction.
V. Manca *(University of Pisa)* Typology and Logical Structure of Natural Languages.
R. Bod *(University of Amsterdam)* Data Oriented Parsing as a General Framework for Stochastic Language Processing.
M. Stefanova & W. ter Stal *(University of Sofia / University of Twente)* A Comparison of ALE and PATR: Practical Experiences.
J.P.M. de Vreught *(University of Delft)* A Practical Comparison between Parallel Tabular Recognizers.
M. Verlinden *(University of Twente)* Head-Corner Parsing of Unification Grammars: A Case Study.
M.-J. Nederhof *(University of Nijmegen)* A Multi-Disciplinary Approach to a Parsing Algorithm.
Th. Strmer *(University of Saarbrcken)* Semantic-Oriented Chart Parsing with Defaults.
G. Satta *(University of Venice)* The Parsing Problem for Tree-Adjoining Grammars.
F. Barthlemy *(University of Lisbon)* A Single Formalism for a Wide Range of Parsers for DCGs.
E. Csuhaj-Varj & R. Abo-Alez *(Hungarian Academy of Sciences, Budapest)* Multi-Agent Systems in Natural Language Processing.
C. Cremers *(University of Leiden)* Coordination as a Parsing Problem.
M. Wirn *(University of Saarbrcken)* Bounded Incremental Parsing.
V. Kubon & M. Platek *(Charles University, Prague)* Robust Parsing and Grammar Checking of Free Word Order Languages.
V. Srinivasan *(University of Mainz)* Punctuation and Parsing of Real-World Texts.
T.G. Vosse *(University of Leiden)* Robust GLR Parsing for Grammar-Based Spelling Correction.

The seventh workshop in the series took place on 15 and 16 june 1994. It was devoted to the topic ”Computer-Assisted Language Learning” (CALL). The aim was to present both the state of the art in CALL and the new perspectives in the research and development of software that is meant to be used in a language curriculum. By the mix of themes addressed in the papers

and demonstrations, we hoped to bring about the exchange of ideas between people of various backgrounds.

Proceedings Twente Workshop on Language Technology 7 (TWLT 7)
Computer-Assisted Language Learning
Eds. L. Appelo, F.M.G. de Jong, 133 pages.

Preface and Contents

L.Appelo, F.M.G. de Jong *(IPO / University of Twente)* Computer-Assisted Language Learning: Prolegomena

M. van Bodegom *(Eurolinguist Language House, Nijmegen, The Netherlands)* Eurolinguist test: An adaptive testing system.

B. Cartigny *(Escape, Tilburg, The Netherlands)* Discatex CD-ROM XA.

H.Altay Guvenir, K. Oflazer *(Bilkent University, Ankara)* Using a Corpus for Teaching Turkish Morphology.

H. Hamburger *(GMU, Washington, USA)* Viewpoint Abstraction: a Key to Conversational Learning.

J. Jaspers, G. Kanselaar, W. Kok *(University of Utrecht, The Netherlands)* Learning English with It's English.

G. Kempen, A. Dijkstra *(University of Leiden, The Netherlands)* Towards an integrated system for spelling, grammar and writing instruction.

F. Kronenberg, A. Krueger, P. Ludewig *(University of Osnabruek, Germany)* Contextual vocabulary learning with CAVOL.

S. Lobbe *(Rotterdam Polytechnic Informatica Centrum, The Netherlands)* Teachers, Students and IT: how to get teachers to integrate IT into the (language) curriculum.

J. Rous, L. Appelo *(Institute for Perception Research, Eindhoven, The Netherlands)* APPEAL: Interactive language learning in a multimedia environment.

B. Salverda *(SLO, Enschede, The Netherlands)* Developing a Multimedia Course for Learning Dutch as a Second Language.

C. Schwind *(Universite de Marseille, France)* Error analysis and explanation in knowledge based language tutoring.

J. Thompson *(CTI, Hull, United Kingdom/EUROCALL)* TELL into the mainstream curriculum.

M. Zock *(Limsi, Paris, France)* Language in action, or learning a language by whatching how it works.

The eighth workshop in the series took place on 1 and 2 December 1994. It was devoted to speech, the integration of speech and natural language processing, and the application of this integration in natural language interfaces. The program emphasized research of interest for the themes in the framework of the Dutch NWO programme on Speech and Natural Language that started in 1994.

Proceedings Twente Workshop on Language Technology 8 (TWLT 8)
Speech and Language Engineering
Eds. L. Boves & A. Nijholt, 176 pages.

Preface and Contents

The ninth workshop in the series took place on 9 June 1995. It was devoted to empirical methods in the analysis of dialogues, and the use of corpora of dialogues in building dialogue systems. The aim was to discuss the methods of corpus analysis, as well as results of corpus analysis and the application of such results.

Proceedings Twente Workshop on Language Technology 9 (TWLT 9)

Corpus-based Approaches to Dialogue Modelling

Eds. J.A. Andernach, S.P. van de Burgt & G.F. van der Hoeven, 124 pages.

J.H. Connolly, A.A. Clarke, S.W. Garner & H.K. Palmn *(Loughborough University of Technology, UK)* Clause-internal structure in spoken dialogue.

J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon & A. Anderson *(HCRC, Edinburgh, UK)* The coding of dialogue structure in a corpus.

J. Alexandersson & N. Reithinger *(DFKI, Saarbrcken, Germany)* Designing the dialogue component in a speech translation system - a corpus-based approach.

H. Aust & M. Oerder *(Philips, Aachen, Germany)* Dialogue control in automatic inquiry systems.

M. Rats *(ITK, Tilburg, the Netherlands)* Referring to topics - a corpus-based study.

H. Dybkjr, L. Dybkjr & N.O. Bernsen *(Centre for Cognitive Science, Roskilde, Denmark)* Design, formalization and evaluation of spoken language dialogue.

D.G. Novick & B. Hansen *(Oregon Graduate Institute of Science and Technology, Portland, USA)* Mutuality strategies for reference in task-oriented dialogue.

N. Fraser *(Vocalis Ltd, Cambridge, UK)* Messy data, what can we learn from it?

J.A. Andernach *(University of Twente, Enschede, the Netherlands)* Predicting and interpreting speech acts in a theatre information and booking system.

The tenth workshop in the series took place on 6-8 December 1995. This workshop was organized in the framework provided by the Algebraic Methodology and Software Technology movement (AMAST). It focussed on algebraic methods in formal languages, programming langauges and natural languages. Its aim was to bring together those researchers on formal language theory, programming language theory and natural language description theory, that have a common interest in the use of algebraic methods to describe syntactic, semantic and pragmatic properties of language.

Proceedings Twente Workshop on Language Technology 10 (TWLT 10)

Algebraic Methods in Language Processing

Eds. A. Nijholt, G. Scollo & R. Steetskamp, 263 pages.

Vladimir A. Fomichov *(Moscow, R)* A Variant of a Universal Metagrammar of Conceptual Structures. Algebraic Systems of Conceptual Syntax.

Theo M.V. Jansen *(Amsterdam, NL)* The Method of ROSETTA, Natural Language Translation Using Algebras.

C.M. Martn-Vide, J. Miquel-Verges & Gh. Paun *(Tarragona, E)* Contextual Grammars with Depth-First Derivation.

Pl Dmsi & Jrgen Duske *(Kussuth University H, University of Hannover, G)* Subword Membership Problem for Linear Indexed Languages.

C. Rico Perez & J.S. Granda *(Madrid, E)* Algebraic Methods for Anaphora Resolution.

Vincenzo Manca *(Pisa, I)* A Logical Formalism for Intergrammatical Representation.

The eleventh workshop in the series took place on 19-21 June 1996. It focussed on the task of dialogue management in natural-language processing systems. The aim was to discuss advances in dialogue management strategies and design methods. During the workshop, there was a separate session concerned with evalutation methods.

Proceedings Twente Workshop on Language Technology 11 (TWLT 11)

Dialogue Management in Natural Language Systems

Eds. S. LuperFoy, A. Nijholt and G. Veldhuijzen van Zanten, 228 pages.

Preface and Contents

David R. Traum *(Universit de Genve, CH)* Conversational Agency: The TRAINS-93 Dialogue Manager.

Scott McGlashan *(SICS, SW)* Towards Multimodal Dialogue Management.

Pierre Nugues, Christophe Godreaux, Pierre-Olivier and Frdric Revolta *(GREYC, F)* A Conversational Agent to Navigate in Virtual Worlds.

Anne Vilnat *(LIMSI-CNRS, F)* Which Processes to Manage Human-Machine Dialogue?

Susann LuperFoy *(MITRE, USA)* Tutoring versus Training: A Mediating Dialogue Manager for Spoken Language Systems.

David G. Novick & Stephen Sutton *(Portland, USA)* Building on Experience: Managing Spoken Interaction through Library Subdialogues.

Latifa Taleb *(INRIA, F)* Communicational Deviation in Finalized InformativeDialogue Management.

Robbert-Jan Beun *(IPO, NL)* Speech Act Generation in Cooperative Dialogue.

Gert Veldhuijzen van Zanten *(IPO, NL)* Pragmatic Interpretation and Dialogue Management in Spoken-Language Systems.

Joris Hulstijn, Ren Steetskamp, Hugo ter Doest, Anton Nijholt & Stan van de Burgt *(University of Twente, NL & KPN Research, NL)* Topics in SCHISMA Dialogues.

Gavin Churcher, Clive Souter & Eric S. Atwell *(Leeds University, UK)* Dialogues in Air Traffic Control

Elisabeth Maier *(DFKI, D)* Context Construction as Subtask of Dialogue Processing – the VERBMOBIL Case.

Anders Baekgaard *(CPK, DK)* Dialogue Management in a Generic Dialogue System.

Wayne Ward *(Carnegie Mellon University, USA)* Dialog Management in the CMU Spoken Language Systems Toolkit.

Wieland Eckert *(University of Erlangen, D)* Understanding of Spontaneous Utterances in Human-Machine-Dialog.

Jan Alexandersson *(DFKI, D)* Some Ideas for the Automatic Acquisition of Dialogue Structure.

Kristiina Jokinen *(Nara Institute of Science and Technology, JP)* Cooperative Response Planning in CDM: Reasoning about Communicative Strategies.

Elizabeth Hinkelman *(Kurzweil Applied Science, USA)* Dialogue Grounding for Speech Recognition Systems.

Jennifer Chu-Carroll *(University of Delaware, USA)* Response Generation in Collaborative Dialogue Interactions.

Harry Bunt *(Tilburg University, NL)* Interaction Management Functions and Context Representation Requirements.

Peter Wyard & Sandra Williams *(BT, GB)* Dialogue Management in a Mixed-Initiative, Cooperative, Spoken Language System.

Rolf Carlson *(KTH, SW)* The Dialog Component in the Waxholm System.

Laila Dybkjr, Niels Ole Bernsen & Hans Dybkjaer *(Roskilde University, DK)* Evaluation of Spoken Dialogue Systems.

Vincenzo Manca *(Pisa, I)* A Logical Formalism for Intergrammatical Representation.

TWLT 12 took place on 11-14 September 1996. It focussed on 'computational humor' and in particular on verbal humor. TWLT12 consisted of a symposium (Marvin Minsky, Douglas Hofstadter, John Allen Paulos, Hugo Brandt Corstius, Oliviero Stock and Gerrit Krol as main speakers), an essay contest for computer science students, two panels, a seminar organized by Salvatore Attardo and Wladyslaw Chlopicki and a two-day workshop (Automatic interpretation and Generation of Verbal Humor) with a mix of invited papers and papers obtained from a Call for Papers.

Proceedings Twente Workshop on Language Technology 12 (TWLT 12)

Computational Humor: Automatic Interpretation and Generation of Verbal Humor

Eds. J. Hulstijn and A. Nijholt, 208 pages.

Preface and Contents

Oliviero Stock 'Password Swordfish': Verbal Humor in the Interface.

Victor Raskin Computer Implementation of the General Theory of Verbal Humor.

Akira Ito & Osamu Takizawa Why do People use Irony? - The Pragmatics of Irony Usage.

Akira Utsumi. Implicit Display Theory of Verbal Irony: Towards a Computational Model of Irony.

Osamu Takizawa, Masuzo Yanagida, Akira Ito & Hitoshi Isahara On Computational Processing of Rhetorical Expressions - Puns, Ironies and Tautologies.

Carmen Curc Relevance Theory and Humorous Interpretations.

Ephraim Nissan From ALIBI to COLOMBUS. The Long March to Self-aware Computational Models of Humor.

Salvatore Attardo Humor Theory beyond Jokes: The Treatment of Humorous Texts at Large.

Bruce Katz A Neural Invariant of Humour.

E. Judith Weiner Why is a Riddle not Like a Metaphor?

Tone Veale No Laughing Matter: The Cognitive Structure of Humour, Metaphor and Creativity.

Tony Veale & Mark Keane Bad Vibes Catastrophes of Goal Activation in the Appreciation of Disparagement Humour and General Poor Taste.

Kim Binsted & Graeme Ritchie Speculations on Story Pun.

Dan Loehr An Integration of a Pun Generator with a Natural Language Robot.

Cameron Shelley, Toby Donaldson & Kim Parsons Humorous Analogy: Modeling 'The Devils Dictionary'.

Michal Ephratt More on Humor Act: What Sort of Speech Act is the Joke?

TWLT 13 took place on 13-15 May 1998. It was the follow-up of the Mundial workshop, that took place in Mnchen in 1997. Both the Mundial workshop as TWLT13 focussed on the formal semantics and pragmatics of dialogues. In addition to the three-day workshop in Twente, with invited and accepted papers, on 18 May a workshop titled 'Communication and Attitudes' was organized at ILLC/University of Amsterdam.

Proceedings Twente Workshop on Language Technology 13 (TWLT 13)

Formal Semantics and Pragmatics of Dialogue (Twendial'98)

Eds. J. Hulstijn and A. Nijholt, 274 pages.

Preface and Contents

Nicholas Asher Varieties of Discourse Structure in Dialogue

Jonathan Ginzburg Clarifying Utterances

Steve Pulman The TRINDI Project: Some Preliminary Themes

Henk Zeevat Contracts in the Common Ground

John Barnden Uncertain Reasoning About Agents' Beliefs and Reasoning, with special attention to Metaphorical Mental State Reports

Thomas Clermont, Marc Pomplun, Elke Prestin and Hannes Rieser Eye-movement Research and the Investigation of Dialogue Structure

Robin Cooper Mixing Situation Theory and Type Theory to Formalize Information States in Dialogue

Jean-louis Dessalles The Interplay of Desire and Necessity in Dialogue

Wieland Eckert Automatic Evaluation of Dialogue Systems

Jelle Gerbrandy Some Remarks on Distributed Knowledge

Jeroen Groenendijk Questions in Update Semantics

Wolfgang Heydrich Theory of Mutuality (Syntactic Skeleton)

Wolfgang Heydrich, Peter Khnlein and Hannes Rieser A DRT-style Modelling of Agents' Mental States in Discourse

Staffan Larsson Questions Under Discussion and Dialogue Moves

Ian Lewin Formal Design, Verification and Simulation of Multi-Modal Dialogues

Nicolas Maudet & Fabrice Evrard A Generic framework for Dialogue Game Implementation

Soo-Jun Park, Keon-Hoe Cha, Won-Kyung Sung, Do Gyu Song, Hyun-A Lee, Jay Duke Park, Dong-In Park & Jrg Hhle MALBOT: An Intelligent Dialogue Model using User Modeling

Massimo Poesio & David Traum Towards an Axiomatization of Dialogue Acts

Mieke Rats Making DRT Suitable for the Description of Information Exchange in a Dialogue

Robert van Rooy Modal subordination in Questions

Adam Zachary Wyner A Discourse Theory of Manner and Factive Adverbial Modification

Marc Blasband A Simple Semantic Model

TWLT14 was held on 7-8 December 1998. It focussed on the role of human language technology in the indexing and accessing of written and spoken documents, video material and/or images, and on the role of language technology for cross-language retrieval and information extraction. The workshop consisted of a series of accepted papers.

Proceedings Twente Workshop on Language Technology 14 (TWLT 14)

Language Technology in Multimedia Information Retrieval

Eds. D. Hiemstra, F.M.G. de Jong and K. Netter, 194 pages.

Marten den Uyl, Ed S. Tan, Heimo Mller & Peter Uray *(SMR Amsterdam, Vrije Universiteit Amsterdam, Joanneum Research)* Towards Automatic Indexing and Retrieval of Video Content: the VICAR system

Anton Nijholt *(University of Twente)* Access, Exploration and Visualization of Interest Communities: The VMC Case Study (in Progress)

Joanne Capstick, Abdel Kader Diagne, Gregor Erbach & Hans Uszkoreit *(DFKI, Saarbrcken)* MULINEX: Multilingual Web Search and Navigation

Klaus Netter & Franciska de Jong *(DFKI, Saarbrcken and University of Twente)* OLIVE: speech based video retrieval

Franciska de Jong *(University of Twente)* Twenty-One: a baseline for multilingual multimedia

---

TWLT15 was held on 19-21 May 1999. It focussed on the interactions in Virtual World. Contributions were invited on theoretical, empirical, computational, experimental, anthropological or philosophical approaches to VR environments. Invited talks were given by Russell Eames *(Microsoft)*, Lewis Johnson (USC), James Lester (North Carolina State University), Pierre Nugues (ISMRA-Caen) and Stephan Matsuba (VRML Dream Company)

Proceedings Twente Workshop on Language Technology 15 (TWLT 15)

Interactions in Virtual Worlds

Eds. Anton Nijholt, Olaf Donk and Betsy van Dijk, 240 pages.

Rainer Kuhn & Sigrun Gujonsdottir *(University of Karlsruhe)* Virtual Campus Project – A Framework for a 3D Multimedia Educational Environment

James C. Lester *(North Carolina State University)* Natural Language Generation in Multimodal Learning Environments

Stephen N. Matsuba *(The VRML Dream Company, Vancouver)* Lifelike Agents and 3D Animated Explanation Generation Speaking Spaces: Virtual Reality, Artificial Intelligence and the Construction of Meaning

Pierre Nugues *(ISMRA-Caen)* Verbal and Written Interaction in Virtual Worlds – Some application examples

Anton Nijholt *(University of Twente)* The Twente Virtual Theatre Environment: Agents and Interactions

Sandy Ressler, Brian Antonishek, Qiming Wang, Afzal Godil & Keith Stouffer *(National Institute of Standards and Technology)* When Worlds Collide –Interactions between the Virtual and the Real

Lakshmi Sastry & D. R. S. Boyd *(Rutherford Appleton Laboratory)* EISCAT Virtual Reality Training Simulation: A Study on Usability and Effectiveness

Frank Schaap *(University of Amsterdam)* "Males say 'blue,' females say 'aqua,' 'sapphire,' and 'dark navy'" The Importance of Gender in Computer-Mediated Communication

Boris van Schooten, Olaf Donk & Job Zwiers *(University of Twente)* Modelling Interaction in Virtual Environments using Process Algebra

Martijn J. Schuemie & Charles A. P. G. van der Mast *(Delft University of Technology)* Presence: Interacting in Virtual Reality?

Jarke J. van Wijk, Bauke de Vries & Cornelius W. A. M. van Overveld *(Eindhoven University of Technology)* Towards an Understanding 3D Virtual Reality Architectural Design System

Peter J. Wyard & Gavin E. Churcher *(BT Laboratories, Ipswich, Suffolk)* Spoken Language Interaction with a Virtual World in the MUeSLI Multimodal 3D Retail System

J. M. Zheng, K. W. Chan & I. Gibson *(University of Hong Kong)* Real Time Gesture Based 3D Graphics User Interface for CAD Modelling System

---

TWLT16/AMiLP 2000 is the second AMAST workshop on Algebraic Methods in Language Processing. Like its predecessor, organized in 1995 at the University of Twente in the TWLT series, papers were presented on formal language theory, programming language theory and natural language theory. A common theme in these papers was the use of mathematics, in particular the use of an algebraic approach. AMiLP 2000 was held in Iowa City, Iowa, USA, from May 20–22 2000, just before the AMAST 2000 conference.

Proceedings Twente Workshop on Language Technology 16 (TWLT 16)

Algebraic Methods in Language Processing (AMiLP2000)

Eds. D. Heylen, A. Nijholt, and G. Scollo, 274 pages.

Preface and Contents

Peter R.J. Asveld *(University of Twente)* Algebraic Aspects of Families of Fuzzy Languages

P. Boullier *(INRIA-Rocquencourt)* 'Counting' with Range Concatenation Grammars

D. Cantone, A. Formisano, E.G. Omodreo & C.G. Zarbu *(University of Catania, University L'Aiquila & University of Perugia)* Compiling Dyadic First-Order Specifications into Map Algebra

---

TWLT 17, *Interacting Agents*, was co-organised with the Centre for Evolutionary Language Engineering (CELE) and as such is also the first workshop in the CELE Workshops on Evolutionary Language Engineering (CEvoLE) series. Together with TWLT18/CEvoLE2 these workshops are jointly titled "Learning to Behave". The workshops investigate human-agent interaction and knowledge both on the level of agents communicating with the external environment and on the level of the internal agent processes that guide the modelling and understanding of the external sensory input in the brain.

TWLT 17 focussed on the interaction of an agent with the environment. This covers many topics such as the use of conversational strategies, natural language and non-verbal communication, turn-taking, protocols, cross-media references, the effect of context, affect and emotion in agents. A special session is devoted to theatre applications. TWLT 17 was held in Enschede from October 18–20 2000.

Proceedings Twente Workshop on Language Technology 17 (TWLT 17)

Learning to Behave, Workshop I: Interacting Agents

Eds. A. Nijholt, D. Heylen and K. Jokinen, 205 pages.

TWLT 18, *Internalising Knowledge*, was co-organised with the Centre for Evolutionary Language Engineering (CELE) and as such is also the first workshop in the CELE Workshops on Evolutionary Language Engineering (CEvoLE) series. Together with TWLT17/CEvoLE1 these workshops are jointly titled "Learning to Behave". The workshops investigate human-agent interaction and knowledge both on the level of agents communicating with the external environment and on the level of the internal agent processes that guide the modelling and understanding of the external sensory input in the brain.

TWLT 18 focussed on internal aspects of learning and interaction: computation in brain-like systens. The goal is to investigate cognitive models for information processing and coordination, especially how symbolic processing, conceptualising and language learning take place in neural models.

TWLT 18 was held in Ieper (Belgium) from November 22-24 2000.

Proceedings Twente Workshop on Language Technology 18 (TWLT 18)
Learning to Behave, Workshop II: Internalising Knowledge

Eds. K. Jokinen, D. Heylen and A. Nijholt, 221 pages.

The proceedings of the workshops can be ordered from Leerstoel TKI, Department of Computer Science, University of Twente, P.O. Box 217, NL-7500 AE Enschede, The Netherlands. E-mail orders are possible: bijron@cs.utwente.nl. Each of the proceedings costs NLG. 50,=.