Christian Kop Günther Fliedl
Heinrich C. Mayr Elisabeth Métais (Eds.)

LNCS 3999

Natural Language Processing and Information Systems

11th International Conference on Applications of Natural Language to Information Systems, NLDB 2006
Klagenfurt, Austria, May/June 2006, Proceedings

-53
.9

Springer

TP18-53
A652.9
2006

Christian Kop Günther Fliedl
Heinrich C. Mayr Elisabeth Métais (Eds.)

Natural Language Processing and Information Systems

11th International Conference on Applications
of Natural Language to Information Systems, NLDB 2006
Klagenfurt, Austria, May 31 – June 2, 2006
Proceedings

华南理工大学
图书馆藏书

E200603604

Springer

Volume Editors

Christian Kop
Günther Fliedl
Heinrich C. Mayr
Alpen-Adria Universität Klagenfurt
Institute of Business Informatics and Application Systems
Klagenfurt, Austria
E-mail: {chris, fliedl, mayr}@ifit.uni-klu.ac.at

Elisabeth Métais
CNAM, Chaire d'Informatique d'Entreprise
292 rue Saint-Martin, 75141 Paris, France
E-mail: metais@cnam.fr

Library of Congress Control Number: 2006926265

CR Subject Classification (1998): H.2, H.3, I.2, F.3-4, H.4, C.2

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-540-34616-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-34616-6 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2006
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11765448 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3999

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison
Lancaster University, UK

Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler
University of Surrey, Guildford, UK

Jon M. Kleinberg
Cornell University, Ithaca, NY, USA

Friedemann Mattern
ETH Zurich, Switzerland

John C. Mitchell
Stanford University, CA, USA

Moni Naor
Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz
University of Bern, Switzerland

C. Pandu Rangan
Indian Institute of Technology, Madras, India

Bernhard Steffen
University of Dortmund, Germany

Madhu Sudan
Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos
University of California, Los Angeles, CA, USA

Doug Tygar
University of California, Berkeley, CA, USA

Moshe Y. Vardi
Rice University, Houston, TX, USA

Gerhard Weikum
Max-Planck Institute of Computer Science, Saarbruecken, Germany

Lecture Notes in Computer Science

For information about Vols. 1–3906

please contact your bookseller or Springer

Vol. 4011: Y. Sure, J. Domingue (Eds.), The Semantic Web: Research and Applications. XVIII, 723 pages. 2006.

Vol. 4007: C. Àlvarez, M. Serna (Eds.), Experimental and Efficient Algorithms. XI, 329 pages. 2006.

Vol. 4004: S. Vaudenay (Ed.), Advances in Cryptology - EUROCRYPT 2006. XIV, 613 pages. 2006.

Vol. 4003: Y. Koucheryavy, J. Harju, V.B. Iversen (Eds.), Next Generation Teletraffic and Wired/Wireless Advanced Networking. XVI, 582 pages. 2006.

Vol. 3999: C. Kop, G. Fliedl, H.C. Mayr, E. Métais (Eds.), Natural Language Processing and Information Systems. XIII, 227 pages. 2006.

Vol. 3998: T. Calamoneri, I. Finocchi, G.F. Italiano (Eds.), Algorithms and Complexity. XII, 394 pages. 2006.

Vol. 3994: V.N. Alexandrov, G.D. van Albada, P.M.A. Sloot, J. Dongarra (Eds.), Computational Science – ICCS 2006, Part IV. XXXV, 1096 pages. 2006.

Vol. 3993: V.N. Alexandrov, G.D. van Albada, P.M.A. Sloot, J. Dongarra (Eds.), Computational Science – ICCS 2006, Part III. XXXVI, 1136 pages. 2006.

Vol. 3992: V.N. Alexandrov, G.D. van Albada, P.M.A. Sloot, J. Dongarra (Eds.), Computational Science – ICCS 2006, Part II. XXXV, 1122 pages. 2006.

Vol. 3991: V.N. Alexandrov, G.D. van Albada, P.M.A. Sloot, J. Dongarra (Eds.), Computational Science – ICCS 2006, Part I. LXXXI, 1096 pages. 2006.

Vol. 3990: J. C. Beck, B.M. Smith (Eds.), Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems. X, 301 pages. 2006.

Vol. 3987: M. Hazas, J. Krumm, T. Strang (Eds.), Location- and Context-Awareness. X, 289 pages. 2006.

Vol. 3986: K. Stølen, W.H. Winsborough, F. Martinelli, F. Massacci (Eds.), Trust Management. XIV, 474 pages. 2006.

Vol. 3984: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), Computational Science and Its Applications - ICCSA 2006, Part V. XXV, 1045 pages. 2006.

Vol. 3983: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), Computational Science and Its Applications - ICCSA 2006, Part IV. XXVI, 1191 pages. 2006.

Vol. 3982: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), Computational Science and Its Applications - ICCSA 2006, Part III. XXV, 1243 pages. 2006.

Vol. 3981: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), Computational Science and Its Applications - ICCSA 2006, Part II. XXVI, 1255 pages. 2006.

Vol. 3980: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), Computational Science and Its Applications - ICCSA 2006, Part I. LXXV, 1199 pages. 2006.

Vol. 3979: T.S. Huang, N. Sebe, M.S. Lew, V. Pavlović, M. Kölsch, A. Galata, B. Kisačanin (Eds.), Computer Vision in Human-Computer Interaction. XII, 121 pages. 2006.

Vol. 3978: B. Hnich, M. Carlsson, F. Fages, F. Rossi (Eds.), Recent Advances in Constraints. VIII, 179 pages. 2006. (Sublibrary LNAI).

Vol. 3976: F. Boavida, T. Plagemann, B. Stiller, C. Westphal, E. Monteiro (Eds.), Networking 2006. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems. XXVI, 1276 pages. 2006.

Vol. 3975: S. Mehrotra, D.D. Zeng, H. Chen, B. Thuraisingham, F.-Y. Wang (Eds.), Intelligence and Security Informatics. XXII, 772 pages. 2006.

Vol. 3973: J. Wang, Z. Yi, J.M. Zurada, B.-L. Lu, H. Yin (Eds.), Advances in Neural Networks - ISNN 2006, Part III. XXIX, 1402 pages. 2006.

Vol. 3972: J. Wang, Z. Yi, J.M. Zurada, B.-L. Lu, H. Yin (Eds.), Advances in Neural Networks - ISNN 2006, Part II. XXVII, 1444 pages. 2006.

Vol. 3971: J. Wang, Z. Yi, J.M. Zurada, B.-L. Lu, H. Yin (Eds.), Advances in Neural Networks - ISNN 2006, Part I. LXVII, 1442 pages. 2006.

Vol. 3970: T. Braun, G. Carle, S. Fahmy, Y. Koucheryavy (Eds.), Wired/Wireless Internet Communications. XIV, 350 pages. 2006.

Vol. 3968: K.P. Fishkin, B. Schiele, P. Nixon, A. Quigley (Eds.), Pervasive Computing. XV, 402 pages. 2006.

Vol. 3967: D. Grigoriev, J. Harrison, E.A. Hirsch (Eds.), Computer Science – Theory and Applications. XVI, 684 pages. 2006.

Vol. 3966: Q. Wang, D. Pfahl, D.M. Raffo, P. Wernick (Eds.), Software Process Change. XIV, 356 pages. 2006.

Vol. 3965: M. Bernardo, A. Cimatti (Eds.), Formal Methods for Hardware Verification. VII, 243 pages. 2006.

Vol. 3964: M. Ü. Uyar, A.Y. Duale, M.A. Fecko (Eds.), Testing of Communicating Systems. XI, 373 pages. 2006.

Vol. 3962: W. IJsselsteijn, Y. de Kort, C. Midden, B. Eggen, E. van den Hoven (Eds.), Persuasive Technology. XII, 216 pages. 2006.

Vol. 3960: R. Vieira, P. Quaresma, M.d.G.V. Nunes, N.J. Mamede, C. Oliveira, M.C. Dias (Eds.), Computational Processing of the Portuguese Language. XII, 274 pages. 2006. (Sublibrary LNAI).

此为试读,需要完整PDF请访问： www.ertongbook.com

Vol. 3959: J.-Y. Cai, S. B. Cooper, A. Li (Eds.), Theory and Applications of Models of Computation. XV, 794 pages. 2006.

Vol. 3958: M. Yung, Y. Dodis, A. Kiayias, T. Malkin (Eds.), Public Key Cryptography - PKC 2006. XIV, 543 pages. 2006.

Vol. 3956: G. Barthe, B. Gregoire, M. Huisman, J.-L. Lanet (Eds.), Construction and Analysis of Safe, Secure, and Interoperable Smart Devices. IX, 175 pages. 2006.

Vol. 3955: G. Antoniou, G. Potamias, C. Spyropoulos, D. Plexousakis (Eds.), Advances in Artificial Intelligence. XVII, 611 pages. 2006. (Sublibrary LNAI).

Vol. 3954: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision – ECCV 2006, Part IV. XVII, 613 pages. 2006.

Vol. 3953: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision – ECCV 2006, Part III. XVII, 649 pages. 2006.

Vol. 3952: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision – ECCV 2006, Part II. XVII, 661 pages. 2006.

Vol. 3951: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision – ECCV 2006, Part I. XXXV, 639 pages. 2006.

Vol. 3950: J.P. Müller, F. Zambonelli (Eds.), Agent-Oriented Software Engineering VI. XVI, 249 pages. 2006.

Vol. 3947: Y.-C. Chung, J.E. Moreira (Eds.), Advances in Grid and Pervasive Computing. XXI, 667 pages. 2006.

Vol. 3946: T.R. Roth-Berghofer, S. Schulz, D.B. Leake (Eds.), Modeling and Retrieval of Context. XI, 149 pages. 2006. (Sublibrary LNAI).

Vol. 3945: M. Hagiya, P. Wadler (Eds.), Functional and Logic Programming. X, 295 pages. 2006.

Vol. 3944: J. Quiñonero-Candela, I. Dagan, B. Magnini, F. d'Alché-Buc (Eds.), Machine Learning Challenges. XIII, 462 pages. 2006. (Sublibrary LNAI).

Vol. 3943: N. Guelfi, A. Savidis (Eds.), Rapid Integration of Software Engineering Techniques. X, 289 pages. 2006.

Vol. 3942: Z. Pan, R. Aylett, H. Diener, X. Jin, S. Göbel, L. Li (Eds.), Technologies for E-Learning and Digital Entertainment. XXV, 1396 pages. 2006.

Vol. 3941: S.W. Gilroy, M.D. Harrison (Eds.), Interactive Systems. XI, 267 pages. 2006.

Vol. 3940: C. Saunders, M. Grobelnik, S. Gunn, J. Shawe-Taylor (Eds.), Subspace, Latent Structure and Feature Selection. X, 209 pages. 2006.

Vol. 3939: C. Priami, L. Cardelli, S. Emmott (Eds.), Transactions on Computational Systems Biology IV. VII, 141 pages. 2006. (Sublibrary LNBI).

Vol. 3936: M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, A. Yavlinsky (Eds.), Advances in Information Retrieval. XIX, 584 pages. 2006.

Vol. 3935: D. Won, S. Kim (Eds.), Information Security and Cryptology - ICISC 2005. XIV, 458 pages. 2006.

Vol. 3934: J.A. Clark, R.F. Paige, F.A. C. Polack, P.J. Brooke (Eds.), Security in Pervasive Computing. X, 243 pages. 2006.

Vol. 3933: F. Bonchi, J.-F. Boulicaut (Eds.), Knowledge Discovery in Inductive Databases. VIII, 251 pages. 2006.

Vol. 3931: B. Apolloni, M. Marinaro, G. Nicosia, R. Tagliaferri (Eds.), Neural Nets. XIII, 370 pages. 2006.

Vol. 3930: D.S. Yeung, Z.-Q. Liu, X.-Z. Wang, H. Yan (Eds.), Advances in Machine Learning and Cybernetics. XXI, 1110 pages. 2006. (Sublibrary LNAI).

Vol. 3929: W. MacCaull, M. Winter, I. Düntsch (Eds.), Relational Methods in Computer Science. VIII, 263 pages. 2006.

Vol. 3928: J. Domingo-Ferrer, J. Posegga, D. Schreckling (Eds.), Smart Card Research and Advanced Applications. XI, 359 pages. 2006.

Vol. 3927: J. Hespanha, A. Tiwari (Eds.), Hybrid Systems: Computation and Control. XII, 584 pages. 2006.

Vol. 3925: A. Valmari (Ed.), Model Checking Software. X, 307 pages. 2006.

Vol. 3924: P. Sestoft (Ed.), Programming Languages and Systems. XII, 343 pages. 2006.

Vol. 3923: A. Mycroft, A. Zeller (Eds.), Compiler Construction. XIII, 277 pages. 2006.

Vol. 3922: L. Baresi, R. Heckel (Eds.), Fundamental Approaches to Software Engineering. XIII, 427 pages. 2006.

Vol. 3921: L. Aceto, A. Ingólfsdóttir (Eds.), Foundations of Software Science and Computation Structures. XV, 447 pages. 2006.

Vol. 3920: H. Hermanns, J. Palsberg (Eds.), Tools and Algorithms for the Construction and Analysis of Systems. XIV, 506 pages. 2006.

Vol. 3918: W.K. Ng, M. Kitsuregawa, J. Li, K. Chang (Eds.), Advances in Knowledge Discovery and Data Mining. XXIV, 879 pages. 2006. (Sublibrary LNAI).

Vol. 3917: H. Chen, F.-Y. Wang, C.C. Yang, D. Zeng, M. Chau, K. Chang (Eds.), Intelligence and Security Informatics. XII, 186 pages. 2006.

Vol. 3916: J. Li, Q. Yang, A.-H. Tan (Eds.), Data Mining for Biomedical Applications. VIII, 155 pages. 2006. (Sublibrary LNBI).

Vol. 3915: R. Nayak, M.J. Zaki (Eds.), Knowledge Discovery from XML Documents. VIII, 105 pages. 2006.

Vol. 3914: A. Garcia, R. Choren, C. Lucena, P. Giorgini, T. Holvoet, A. Romanovsky (Eds.), Software Engineering for Multi-Agent Systems IV. XIV, 255 pages. 2006.

Vol. 3911: R. Wyrzykowski, J. Dongarra, N. Meyer, J. Waśniewski (Eds.), Parallel Processing and Applied Mathematics. XXIII, 1126 pages. 2006.

Vol. 3910: S.A. Brueckner, G.D.M. Serugendo, D. Hales, F. Zambonelli (Eds.), Engineering Self-Organising Systems. XII, 245 pages. 2006. (Sublibrary LNAI).

Vol. 3909: A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 612 pages. 2006. (Sublibrary LNBI).

Vol. 3908: A. Bui, M. Bui, T. Böhme, H. Unger (Eds.), Innovative Internet Community Systems. VIII, 207 pages. 2006.

Vol. 3907: F. Rothlauf, J. Branke, S. Cagnoni, E. Costa, C. Cotta, R. Drechsler, E. Lutton, P. Machado, J.H. Moore, J. Romero, G.D. Smith, G. Squillero, H. Takagi (Eds.), Applications of Evolutionary Computing. XXIV, 813 pages. 2006.

¥359.00元

Preface

Information systems and natural language processing are fundamental fields of research and development in informatics. The combination of both is an exciting and future-oriented field which has been addressed by the NLDB conference series since 1995. There are still many open research questions but also an increasing number of interesting solutions and approaches.

NLDB 2006 with its high-quality contributions tersely reflected the current discussion and research: natural language and/or ontology-based information retrieval, question-answering methods, dialog processing, query processing as well as ontology- and concept creation from natural language. Some papers presented the newest methods for parsing, entity recognition and language identification which are important for many of the topics mentioned before. In particular, 53 papers were submitted by authors from 14 nations. From these contributions, the Program Committee, based on 3 peer reviews for each paper, selected 17 full and 5 short papers, thus coming up with an overall acceptance rate of 32% (41% including short papers).

Many persons contributed to making NLDB 2006 a success. First we thank all authors for their valuable contributions. Secondly, we thank all members of the Program Committee for their detailed reviews and discussion. Furthermore we thank the following people for their substantial organizational collaboration: Kerstin Jörgl, who did a lot of work to compose these proceedings, our Conference Secretary, Christine Seger, Stefan Ellersdorfer for his technical support, and Jürgen Vöhringer and Christian Winkler, who provided additional last-minute reviews.

This year, NLDB was a part of a multi-conference event on Information Systems: UNISCON — United Information Systems Conference. Thus, participants could get into scientific contact with experts from more technical (ISTA 2006) or more business-oriented (BIS 2006) fields. In either case, they profited from UNISCON's organizational environment. We, therefore, express our thanks also to the UNISCON organization team: Markus Adam, Jörg Kerschbaumer and all the students who supported the participants during the NLDB 2006 conference.

March 2006

Christian Kop
Günther Fliedl
Heinrich C. Mayr
Eliabeth Métais

Nadira Lammari	CEDRIC/CNAM, France
Winfried Lenders	Universität Bonn, Germany
Jana Lewerenz	sd&m Düsseldorf, Germany
Stephen Liddle	Brigham Young University, USA
Deryle Lonsdale	Brigham Young Uinversity, USA
Robert Luk	Hong Kong Polytechnic University, Hong Kong
Heinrich C. Mayr	University of Klagenfurt, Austria
Elisabeth Métais	CEDRIC/CNAM , France
Farid Meziane	Salford University, UK
Luisa Mich	University of Trento, Italy
Diego Mollá Aliod	Macquarie University, Australia
Andrés Montoyo	Universidad de Alicante, Spain
Ana Maria Moreno	Universidad Politecnica de Madrid, Spain
Rafael Muñoz	Universidad de Alicante, Spain
Günter Neumann	DFKI, Germany
Jian-Yun Nie	Université de Montréal, Canada
Manual Palomar	Universidad de Alicante, Spain
Sandeep Purao	Pennsylvania State University, USA
Odile Piton	Université Paris I Panthéon-Sorbonne, France
Yacine Rezgui	University of Salford, UK
Reind van de Riet	Vrije Universiteit Amsterdam, The Netherlands
Hae-Chang Rim	Korea University, Korea
Veda Storey	Georgia State University, USA
Vijay Sugumaran	Oakland University Rochester, USA
Bernhard Thalheim	Kiel University, Germany
Krishnaprasad Thirunarayan	Wright State University, USA
Juan Carlos Trujillo	Universidad de Alicante, Spain
Luis Alfonso Ureña	Universidad de Jaén, Spain
Sunil Vadera	University of Salford, UK
Panos Vassiliadis	University of Ioannina, Greece
Jürgen Vöhringer	University of Klagenfurt, Austria
Hans Weigand	Tilburg University, The Netherlands
Werner Winiwarter	University of Vienna, Austria
Christian Winkler	University of Klagenfurt, Austria

External Referees

Birger Andersson
Maria Bergholtz
Miguel Ángel García Cumbreras
Theodore Dalamagas
Hiroshi Echizen-ya
Yasutomo Kimura
Nadia Kiyavitskaya
Francisco Javier Ariza López

Organization

Conference Co-chairs

Christian Kop	Alpen-Adria Universität Klagenfurt
Günther Fliedl	Alpen-Adria Universität Klagenfurt
Heinrich C. Mayr	Alpen-Adria Universität Klagenfurt
Elisabeth Métais	Cedric Laboratory CNAM, Paris

Organization and Local Arrangements

Markus Adam	Alpen-Adria Universität Klagenfurt
Stefan Ellersdorfer	Alpen-Adria Universität Klagenfurt
Kerstin Jörgl	Alpen-Adria Universität Klagenfurt
Christine Seger	Alpen-Adria Universität Klagenfurt
Christian Winkler	Alpen-Adria Universität Klagenfurt

Program Committee

Kenji Araki	Hokkaido University, Japan
Akhilesh Bajaj	University of Tulsa, USA
Mokrane Bouzeghoub	PRiSM, Université de Versailles, France
Andrew Burton-Jones	University of British Columbia, Canada
Roger Chiang	University of Cincinnati, USA
Gary A. Coen	Boeing, USA
Isabelle Comyn-Wattiau	CEDRIC/CNAM, France
Antje Düsterhöft	University of Wismar, Germany
Günther Fliedl	Universität Klagenfurt, Austria
Alexander Gelbukh	Instituto Politecnico Nacional, Mexico
Nicola Guarino	CNR, Italy
Jon Atle Gulla	Norwegian University of Science and Technology, Norway
Karin Harbusch	Universität Koblenz-Landau, Germany
Helmut Horacek	Universität des Saarlandes, Germany
Cecil Chua Eng Huang	Nanyang Technological University, Singapore
Paul Johannesson	Stockholm University, Sweden
Zoubida Kedad	PRiSM, Université de Versailles, France
Christian Kop	University of Klagenfurt, Austria
Leila Kosseim	Concordia University, Canada

Borja Navarro
Lluís Padró
Hideyuki Shibuki
Darijus Strasunskas
Stein L. Tomassen
Sonia Vazquez
Chih-Sheng Yang
Nicola Zeni

Organized by:

NLDB was organized by the Institute of Business Informatics and Applications Systems, Alpen-Adria University of Klagenfurt, Austria.

Table of Contents

Concepts Extraction and Ontology

Ontologies and Task Repository Utilization

Query Processing

此为试读,需要完整PDF请访问: www.ertongbook.com

Information Retrieval and Dialog Processing

NLP Techniques

Short Paper Session I

Short Paper Session II

An Automated Multi-component Approach to Extracting Entity Relationships from Database Requirement Specification Documents

Siqing Du and Douglas P. Metzler

University of Pittsburgh, School of Information Sciences, Pittsburgh,
PA, 15260, USA
sid2@pitt.edu, metzler@mail.sis.pitt.edu

Abstract. This paper describes a natural language system that extracts entity relationship diagram components from natural language database design documents. The system is a fully integrated composite of existing, publicly available components including a parser, WordNet and Google web corpus search facilities, and a novel rule-based tuple-extraction process. The system differs from previous approaches in being fully automatic (as opposed to approaches requiring human disambiguation or other interaction) and in providing a higher level of performance than previously reported results.

1 Introduction

In the database design process, designers first elicit natural language requirements specifications from users, then transform the requirements into a formal representation schema. The Entity Relationship Diagram (ERD) model is one of the most popular of these formal representation schemata [1]. It is a graphic way of displaying entity, relationship, and attribute types, which incorporates some of the important semantic information about the real world situation, and it is considered to be a very natural and easy-to-understand way of conceptualizing the structure of a database.

The process of translating natural language Database Requirement Specification (DRS) documents into ERDs is, however, time-consuming, error-prone and costly, and there have been a series of attempts to automate this process, which is relevant both to research in sublanguage processing (since the language of database requirement specifications is more limited than general natural language both in structure and content) and to the general problem of extraction of formal representations from natural language. Chen [2] first explored the correspondence between English sentence structure and that of entity relationship diagrams and proposed eleven rules for translation. These rules were neither complete nor fully accurate and a number of studies have tried to improve on this approach [3-9]. Some of these approaches, e.g., [3] emphasized a dialogue tool that helped elicit the natural language description itself, while at the same time avoiding some of the disambiguation difficulties of the natural language translation process.

Most approaches to this problem have involved extensions of Chen's original approach [2], involving the application of heuristic translation rules to the outputs of a

C. Kop et al. (Eds.): NLDB 2006, LNCS 3999, pp. 1 – 11, 2006.
© Springer-Verlag Berlin Heidelberg 2006

syntactic parser. Most recently Omar *et al.* [4] used such an approach augmented with a confidence level weighting mechanism to achieve a reported performance of 95% recall and 82% precision in terms of entity recognition. They did not however describe the results in terms of the correctness of attachments of entities and relations.

Automated extraction of entity relationship diagrams from natural language DRS is hard due to the lexical, syntactic and semantic ambiguity of natural language. To avoid the need for human input to resolve such ambiguity, it is necessary to go beyond the simple models incorporating syntactic parsing and simple translation rules, to utilize the semantic information required for such disambiguation. For some natural language applications this can require extensive and explicit knowledge bases that are expensive and time consuming to construct and often do not fully cover the necessary knowledge.

This paper describes an integrated multi-component framework for this problem that brings the necessary semantic knowledge to bear using publicly available, relatively weakly structured, but very large scale resources. It is suggested that this architecture, or ones similar to it, will be applicable to a wide set of similar special purpose natural language to formalism translation problems, and perhaps to a wider set of limited scope natural language problems as well.

2 A Model

Most of the previous research on applying natural language processing to database concept modeling employed a model consisting of the application of heuristic translation rules to the outputs of a syntactic parser. This model proved incapable of adequately dealing with issues such as distinguishing entities from attributes, recognizing conjunctive relations and subtle semantic disambiguation. This section describes an enhanced general model incorporating two knowledge sources capable of providing elements of semantic processing. Each of the components can be realized by different choices according to specific domain and application requirements. The specific components utilized in this system, and the advantages of the approaches taken with some of the components will be described in the following section on implementation.

2.1 A Syntactic Parser

This model follows previous research on ERD extraction, as well as most natural language processing work in general, in having an independent syntactic parser, as opposed to incorporating semantic and syntactic considerations in a single formalism. Some criteria for the choice of syntactic parser are obvious, such as accuracy and lexical coverage. Others are perhaps less so. It is desirable to have an extensible lexicon, so that DRS documents written for particular specialized domains can be covered. The parser should be robust in handling incomplete or incorrect structures, and it should allow for user extensions to cover additional domain specific structures. Finally, the output structures should facilitate the types of processing required by the following components of the system.

2.2 Heuristic Translation Rules

As in previous work, translation rules are employed to extract the ERD components from the outputs of the syntactic parser. In this system, the forms of the translation rules are unusually tightly coupled to the format of the outputs of the syntactic parser. Also, because of the use of semantic filters, the translation rules can be allowed to overgenerate to a certain degree, reducing potential problems of undergeneration. Because of the inherent ambiguity of natural language, it is possible to extract multiple ERDs from a given natural description. What is desired is that interpretations that violate semantic constraints are not retained.

2.3 A Lexical Filter

One form of semantic filtering involves a lexical filter which can help in issues such as distinguishing entities from attributes and relations. Although most modern parsers are lexicalized, it can be useful to employ a separate, post-parsing, lexical filter to deal with semantic disambiguation issues concerned with the semantic interpretations of parsing outputs, as opposed to lexical issues concerned with deriving the correct syntactic interpretations. A post-parsing lexical component can be in the form of a dictionary, lexicon, thesaurus, or general ontology. It can also incorporate domain specific modules. Using a separate post-parsing lexical component allows a system designer to employ a wider and less constrained set of approaches to lexical disambiguation and filtering.

2.4 A Semantic Filter

Some classical ambiguity problems such as prepositional phrase attachment and conjunctive structure interpretation require an analysis of the appropriateness or likeliness of relationships among terms. Some sort of semantic analysis of complex, multi-term structures is necessary to handle these issues. The choices here range from very detailed knowledge-based approaches, which are notoriously difficult to provide for large open-ended domains, to shallow quasi-statistical approaches based on the empirical evidence found in large corpora.

3 An Implementation

In this general approach, one could utilize a number of options for each component. In addition to the specific reasons for the choices made which are discussed below, there were several general considerations that guided the choices. The components utilized (with the exception of the translation rules) are powerful off-the-shelf components that lend themselves to integration in a straightforward environment. We utilized the Link Parser and WordNet, which are open source C language resources that can be readily compiled as dynamic libraries, and Google APIs, which are used to access the Google web corpus via SOAP and WSDL standards. All of them were integrated in the .Net environment.

3.1 Link Parser

The parsing component utilizes Link Parser (LP), which is based on Link Grammar, an original theory of English syntax [10]. The parser has a lexicon of about 60,000 word forms. It covers a wide variety of syntactic constructions, including many rare and idiomatic ones. It also can be easily extended [11]. The parser is robust [12]; it is able to skip over portions of the sentence that it cannot understand, and assign some structure to the rest of the sentence. It is able to handle unknown vocabulary, and make intelligent guesses from context and spelling about the syntactic categories of unknown words. It has knowledge of capitalization, numerical expressions, and a variety of punctuation symbols. Link Grammar and LP have been applied to several sublanguages to extract useful information [13-15]. LP is similar to the Constituent Object Parser [16] and Link Grammar is similar to ideas of [17] upon which [16] was loosely based.

The basic idea of Link Grammar (and these related approaches) is to transfer some of the formal complexity of context-free grammar rules to the data structures representing words. Words are structures with left- and right-pointing connectors. Left- and right-pointing connectors can be joined to form a link, and words have rules that determine how the links can be formed. A sentence is accepted as valid when a complete linkage is formed that includes all the words of the sentence. LP can return more than one complete linkage for a single sentence, reflecting the fact that sentences do have valid alternative syntactic interpretations. LP uses probabilistic information on the creation of links to estimate the likelihood of the correctness of individual interpretations of a given sentence, and its performance in that regard is good. Informally, it appeared that the most likely syntactic interpretation of sentences was almost always among the top three or four interpretations returned. The use of the semantic filters in this architecture made this kind of performance adequate for this application.

3.2 Heuristic Translation Rules

The translation rules extract entity-relation connections which are represented as tuples, e.g. <*has company employees*>. For an *n*-ary relationship, there are total *n*+1 objects in a tuple. The first one is the relationship; the following are the entities connected by the relationship. Each of the objects also has some properties. The LP was chosen over similarly powerful parsers because the links correspond relatively directly to the components of entity relation diagrams, as illustrated in Fig. 1. for the sentence " *A company has 3,000 employees*".

In Fig. 1., the "Ss"(subject) and "Op"(object) links in Link Grammar correspond to the two entity-to-relation links in ERD. The "Dsu" and "Dmcn" links provide cardinality information for the relation between the two entities. This basic pattern can be extracted even from a more elaborate and complex version of this sentence, and, in general, such ERD patterns can be extracted from a wide variety of complex sentence forms. The LP employs 107 link types that describe relationships between words, but many of these are not relevant to the style of language used in DRS or are not relevant to the ERD extraction process. The extraction/translation module consists of 50 rules, each of which applies to a single link type. The rules are not strictly declarative, but rather include procedural actions and calls, even to outside modules such as the WordNet component.

The heuristic translation rules we used are link-based, which are more flexible and semantic-richer, compared with syntactic-driven, POS-based approaches in previous research [2, 4, 7]. For each heuristic rule, it is verb-centered. A frame is built for each verb. Then various rules are applied to fill out the values of the slots of the frame,

此为试读，需要完整PDF请访问：www.ertongbook.com