Andrés Montoyo
Rafael Muñoz
Elisabeth Métais (Eds.)

LNCS 3513

Natural Language Processing and Information Systems

10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005 Alicante, Spain, June 2005, Proceedings

Springer

Andrés Montoyo Rafael Muñoz
Elisabeth Métais (Eds.)

Natural Language Processing and Information Systems

10th International Conference on Applications
of Natural Language to Information Systems, NLDB 2005
Alicante, Spain, June 15-17, 2005
Proceedings

Springer

Volume Editors

Andrés Montoyo
Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
Apartado de correos, 99, 03080 Alicante, Spain
E-mail: montoyo@dlsi.ua.es

Rafael Muñoz
Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
Apartado de correos, 99, 03080 Alicante, Spain
E-mail: rafael@dlsi.ua.es

Elisabeth Métais
Cedric Laboratory, CNAM
292 rue Saint-Martin, 75003 Paris, France
E-mail: elsa@cnam.fr

Library of Congress Control Number: Applied for

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

ISSN 0302-9743
ISBN-10 3-540-26031-5 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-26031-8 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11428817 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3513

Preface

NLDB 2005, the 10th International Conference on Applications of Natural Language to Information Systems, was held on June 15–17, 2005 at the University of Alicante, Spain. Since the first NLDB conference in 1995 the main goal has been to provide a forum to discuss and disseminate research on the integration of natural language resources in information system engineering.

The development and convergence of computing, telecommunications and information systems has already led to a revolution in the way that we work, communicate with each other, buy goods and use services, and even in the way that we entertain and educate ourselves. The revolution continues, and one of its results is that large volumes of information will increasingly be held in a form which is more natural for users than the data presentation formats typical of computer systems of the past. Natural language processing (NLP) is crucial in solving these problems, and language technologies will make an indispensable contribution to the success of information systems. We hope that NLDB 2005 was a modest contribution to this goal.

NLDB 2005 contributed to advancing the goals and the high international standing of these conferences, largely due to its Program Committee, composed of renowned researchers in the field of natural language processing and information system engineering. Papers were reviewed by three reviewers from the Program Committee. This clearly contributed to the significant number of papers submitted (95). Twenty-nine were accepted as regular papers, while 18 were accepted as short papers.

We would like to express here our thanks to all the reviewers for their quick and excellent work. We extend these thanks to our invited speakers, Ruslan Mitkov and Branimir Boguraev, for their valuable contribution, which undoubtedly increased the interest in the conference. We are also indebted to a number of individuals for taking care of specific parts of the conference program, specially to Miguel Angel Varó who built and maintained all Web services for the conference.

March 2005

Andres Montoyo
Rafael Muñoz
Elisabeth Métais

Organization

Conference Chairs

Rafael Muñoz (University of Alicante, Spain)
Elisabeth Métais (CEDRIC/CNAM, France)

Program Chair

Andres Montoyo (University of Alicante, Spain)

Organization Committee

Patricio Martínez-Barco (University of Alicante, Spain)
Andres Montoyo (University of Alicante, Spain)
Paloma Moreda (University of Alicante, Spain)
Rafael Muñoz (University of Alicante, Spain)
Maximiliano Saiz (University of Alicante, Spain)
Armando Suárez (University of Alicante, Spain)
Elisa Noguera (University of Alicante, Spain)

Program Committee

Kenji Araki (Hokkaido University, Japan)
Mokrane Bouzeghoub (PRiSM, Université de Versailles, France)
Gary A. Coen (Boeing, USA)
Isabelle Comyn-Wattiau (CEDRIC/CNAM, France)
Günther Fliedl (Universität Klagenfurt, Austria)
Alexander Gelbukh (Instituto Politécnico Nacional, Mexico)
Jon Atle Gulla (Norwegian University of Science and Technology, Norway)
Harmain Harmain (United Arab Emirates University, United Arab Emirates)
Helmut Horacek (Universität des Saarlandes, Germany)
Paul Johannesson (Stockholm University, Sweden)
Zoubida Kedad (PRiSM, Université de Versailles, France)
Nadira Lammari (CEDRIC/CNAM, France)
Jana Lewerenz (sd&m Düsseldorf, Germany)
Robert Luk (Hong Kong Polytechnic University, Hong Kong, China)
Bernardo Magnini (IRST, Italy)
Paul McFetridge (Simon Frazer University, Canada)
Elisabeth Métais (CEDRIC/CNAM, France)
Farid Meziane (Salford University, UK)

Luisa Mich (University of Trento, Italy)
Ruslan Mitkov (University of Wolverhampton, UK)
Ana Maria Moreno (Universidad Politécnica de Madrid, Spain)
Diego Mollá Aliod (Macquarie University, Australia)
Andrés Montoyo (Universidad de Alicante, Spain)
Rafael Muñoz (Universidad de Alicante, Spain)
Jian-Yun Nie (Université de Montréal, Canada)
Manuel Palomar (Universidad de Alicante, Spain)
Odile Piton (Université Paris I Panthéon-Sorbonne, France)
Reind van de Riet (Vrije Universiteit Amsterdam, The Netherlands)
Hae-Chang Rim (Korea University, Korea)
Vijay Sugumaran (Oakland University Rochester, USA)
Veda Storey (Georgia State University, USA)
Bernhard Thalheim (Kiel University, Germany)
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)
Hans Weigand (Tilburg University, The Netherlands)
Werner Winiwarter (University of Vienna, Austria)
Christian Winkler (Universität Klagenfurt, Austria)
Stanislaw Wrycza (University of Gdansk, Poland)

Additional Reviewers

Andrea, Mulloni
Bergholtz, Maria
Isabelle, Comyn-Wattiau
Nadira, Lammari
Del Jesus Diaz, Maria Jose
Echizen-ya, Hiroshi
Evans, Richard
Ferrández, Antonio
García Cumbreras, Miguel Ángel
Ha, Le An
Haddad, Hatem
Kabilan, Vandana
Kozareva, Zortnisa
Llopis, Fernando
Maria, Bergholtz
Martínez-Barco, Patricio
Martin-Valdivia, M.Teresa
McFetridge, Paul
Montejo Ráez, Arturo
Moreda, Paloma
Navarro, Borja
Pekar, Viktor
Peral, Jesús
Prost, Jean-Philippe
Puscasu, Georgiana
Roger, Sandra
Rzepka, Rafal
Sasaoka, Hisayuki
Selima, Besbes
Storey, Veda
Vadera, Sunil
Vazquez, Sonia
Xue, Xiaohui

Lecture Notes in Computer Science

For information about Vols. 1–3407

please contact your bookseller or Springer

Vol. 3532: A. Gómez-Pérez, J. Euzenat (Eds.), The Semantic Web: Research and Applications. XV, 728 pages. 2005.

Vol. 3525: A.E. Abdallah, C.B. Jones, J.W. Sanders (Eds.), Communicating Sequential Processes. XIV, 321 pages. 2005.

Vol. 3518: T.B. Ho, D. Cheung, H. Li (Eds.), Advances in Knowledge Discovery and Data Mining. XXI, 864 pages. 2005. (Subseries LNAI).

Vol. 3517: H.S. Baird, D.P. Lopresti (Eds.), Human Interactive Proofs. IX, 143 pages. 2005.

Vol. 3516: V.S. Sunderam, G.D.v. Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part III. LXIII, 1143 pages. 2005.

Vol. 3515: V.S. Sunderam, G.D.v. Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part II. LXIII, 1101 pages. 2005.

Vol. 3514: V.S. Sunderam, G.D.v. Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part I. LXIII, 1089 pages. 2005.

Vol. 3513: A. Montoyo, R. Muñoz, E. Métais (Eds.), Natural Language Processing and Information Systems. XII, 408 pages. 2005.

Vol. 3510: T. Braun, G. Carle, Y. Koucheryavy, V. Tsaoussidis (Eds.), Wired/Wireless Internet Communications. XIV, 366 pages. 2005.

Vol. 3508: P. Bresciani, P. Giorgini, B. Henderson-Sellers, G. Low, M. Winikoff (Eds.), Agent-Oriented Information Systems II. X, 227 pages. 2005. (Subseries LNAI).

Vol. 3505: V. Gorodetsky, J. Liu, V. A. Skormin (Eds.), Autonomous Intelligent Systems: Agents and Data Mining. XIII, 303 pages. 2005. (Subseries LNAI).

Vol. 3503: S.E. Nikoletseas (Ed.), Experimental and Efficient Algorithms. XV, 624 pages. 2005.

Vol. 3502: F. Khendek, R. Dssouli (Eds.), Testing of Communicating Systems. X, 381 pages. 2005.

Vol. 3501: B. Kégl, G. Lapalme (Eds.), Advances in Artificial Intelligence. XV, 458 pages. 2005. (Subseries LNAI).

Vol. 3500: S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 632 pages. 2005. (Subseries LNBI).

Vol. 3499: A. Pelc, M. Raynal (Eds.), Structural Information and Communication Complexity. X, 323 pages. 2005.

Vol. 3498: J. Wang, X. Liao, Z. Yi (Eds.), Advances in Neural Networks – ISNN 2005, Part III. L, 1077 pages. 2005.

Vol. 3497: J. Wang, X. Liao, Z. Yi (Eds.), Advances in Neural Networks – ISNN 2005, Part II. L, 947 pages. 2005.

Vol. 3496: J. Wang, X. Liao, Z. Yi (Eds.), Advances in Neural Networks – ISNN 2005, Part II. L, 1055 pages. 2005.

Vol. 3495: P. Kantor, G. Muresan, F. Roberts, D.D. Zeng, F.-Y. Wang, H. Chen, R.C. Merkle (Eds.), Intelligence and Security Informatics. XVIII, 674 pages. 2005.

Vol. 3494: R. Cramer (Ed.), Advances in Cryptology – EUROCRYPT 2005. XIV, 576 pages. 2005.

Vol. 3492: P. Blache, E. Stabler, J. Busquets, R. Moot (Eds.), Logical Aspects of Computational Linguistics. X, 363 pages. 2005. (Subseries LNAI).

Vol. 3489: G.T. Heineman, I. Crnkovic, H.W. Schmidt, J.A. Stafford, C. Szyperski, K. Wallnau (Eds.), Component-Based Software Engineering. XI, 358 pages. 2005.

Vol. 3488: M.-S. Hacid, N.V. Murray, Z.W. Raś, S. Tsumoto (Eds.), Foundations of Intelligent Systems. XIII, 700 pages. 2005. (Subseries LNAI).

Vol. 3486: T. Helleseth, D. Sarwate, H.-Y. Song, K. Yang (Eds.), Sequences and Their Applications - SETA 2004. XII, 451 pages. 2005.

Vol. 3483: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part IV. XXVII, 1362 pages. 2005.

Vol. 3482: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part III. LXVI, 1340 pages. 2005.

Vol. 3481: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part II. LXIV, 1316 pages. 2005.

Vol. 3480: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part I. LXV, 1234 pages. 2005.

Vol. 3479: T. Strang, C. Linnhoff-Popien (Eds.), Location- and Context-Awareness. XII, 378 pages. 2005.

Vol. 3478: C. Jermann, A. Neumaier, D. Sam (Eds.), Global Optimization and Constraint Satisfaction. XIII, 193 pages. 2005.

Vol. 3477: P. Herrmann, V. Issarny, S. Shiu (Eds.), Trust Management. XII, 426 pages. 2005.

Vol. 3475: N. Guelfi (Ed.), Rapid Integration of Software Engineering Techniques. X, 145 pages. 2005.

Vol. 3468: H.W. Gellersen, R. Want, A. Schmidt (Eds.), Pervasive Computing. XIII, 347 pages. 2005.

Vol. 3467: J. Giesl (Ed.), Term Rewriting and Applications. XIII, 517 pages. 2005.

Vol. 3465: M. Bernardo, A. Bogliolo (Eds.), Formal Methods for Mobile Computing. VII, 271 pages. 2005.

Vol. 3463: M. Dal Cin, M. Kaâniche, A. Pataricza (Eds.), Dependable Computing - EDCC 2005. XVI, 472 pages. 2005.

Vol. 3462: R. Boutaba, K. Almeroth, R. Puigjaner, S. Shen, J.P. Black (Eds.), NETWORKING 2005. XXX, 1483 pages. 2005.

Vol. 3461: P. Urzyczyn (Ed.), Typed Lambda Calculi and Applications. XI, 433 pages. 2005.

Vol. 3460: Ö. Babaoglu, M. Jelasity, A. Montresor, C. Fetzer, S. Leonardi, A. van Moorsel, M. van Steen (Eds.), Self-star Properties in Complex Information Systems. IX, 447 pages. 2005.

Vol. 3459: R. Kimmel, N.A. Sochen, J. Weickert (Eds.), Scale Space and PDE Methods in Computer Vision. XI, 634 pages. 2005.

Vol. 3458: P. Herrero, M.S. Pérez, V. Robles (Eds.), Scientific Applications of Grid Computing. X, 208 pages. 2005.

Vol. 3456: H. Rust, Operational Semantics for Timed Systems. XII, 223 pages. 2005.

Vol. 3455: H. Treharne, S. King, M. Henson, S. Schneider (Eds.), ZB 2005: Formal Specification and Development in Z and B. XV, 493 pages. 2005.

Vol. 3454: J.-M. Jacquet, G.P. Picco (Eds.), Coordination Models and Languages. X, 299 pages. 2005.

Vol. 3453: L. Zhou, B.C. Ooi, X. Meng (Eds.), Database Systems for Advanced Applications. XXVII, 929 pages. 2005.

Vol. 3452: F. Baader, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XI, 562 pages. 2005. (Subseries LNAI).

Vol. 3450: D. Hutter, M. Ullmann (Eds.), Security in Pervasive Computing. XI, 239 pages. 2005.

Vol. 3449: F. Rothlauf, J. Branke, S. Cagnoni, D.W. Corne, R. Drechsler, Y. Jin, P. Machado, E. Marchiori, J. Romero, G.D. Smith, G. Squillero (Eds.), Applications of Evolutionary Computing. XX, 631 pages. 2005.

Vol. 3448: G.R. Raidl, J. Gottlieb (Eds.), Evolutionary Computation in Combinatorial Optimization. XI, 271 pages. 2005.

Vol. 3447: M. Keijzer, A. Tettamanzi, P. Collet, J.v. Hemert, M. Tomassini (Eds.), Genetic Programming. XIII, 382 pages. 2005.

Vol. 3444: M. Sagiv (Ed.), Programming Languages and Systems. XIII, 439 pages. 2005.

Vol. 3443: R. Bodik (Ed.), Compiler Construction. XI, 305 pages. 2005.

Vol. 3442: M. Cerioli (Ed.), Fundamental Approaches to Software Engineering. XIII, 373 pages. 2005.

Vol. 3441: V. Sassone (Ed.), Foundations of Software Science and Computational Structures. XVIII, 521 pages. 2005.

Vol. 3440: N. Halbwachs, L.D. Zuck (Eds.), Tools and Algorithms for the Construction and Analysis of Systems. XVII, 588 pages. 2005.

Vol. 3439: R.H. Deng, F. Bao, H. Pang, J. Zhou (Eds.), Information Security Practice and Experience. XII, 424 pages. 2005.

Vol. 3437: T. Gschwind, C. Mascolo (Eds.), Software Engineering and Middleware. X, 245 pages. 2005.

Vol. 3436: B. Bouyssounouse, J. Sifakis (Eds.), Embedded Systems Design. XV, 492 pages. 2005.

Vol. 3434: L. Brun, M. Vento (Eds.), Graph-Based Representations in Pattern Recognition. XII, 384 pages. 2005.

Vol. 3433: S. Bhalla (Ed.), Databases in Networked Information Systems. VII, 319 pages. 2005.

Vol. 3432: M. Beigl, P. Lukowicz (Eds.), Systems Aspects in Organic and Pervasive Computing - ARCS 2005. X, 265 pages. 2005.

Vol. 3431: C. Dovrolis (Ed.), Passive and Active Network Measurement. XII, 374 pages. 2005.

Vol. 3429: E. Andres, G. Damiand, P. Lienhardt (Eds.), Discrete Geometry for Computer Imagery. X, 428 pages. 2005.

Vol. 3428: Y.-J. Kwon, A. Bouju, C. Claramunt (Eds.), Web and Wireless Geographical Information Systems. XII, 255 pages. 2005.

Vol. 3427: G. Kotsis, O. Spaniol (Eds.), Wireless Systems and Mobility in Next Generation Internet. VIII, 249 pages. 2005.

Vol. 3423: J.L. Fiadeiro, P.D. Mosses, F. Orejas (Eds.), Recent Trends in Algebraic Development Techniques. VIII, 271 pages. 2005.

Vol. 3422: R.T. Mittermeir (Ed.), From Computer Literacy to Informatics Fundamentals. X, 203 pages. 2005.

Vol. 3421: P. Lorenz, P. Dini (Eds.), Networking - ICN 2005, Part II. XXXV, 1153 pages. 2005.

Vol. 3420: P. Lorenz, P. Dini (Eds.), Networking - ICN 2005, Part I. XXXV, 933 pages. 2005.

Vol. 3419: B. Faltings, A. Petcu, F. Fages, F. Rossi (Eds.), Constraint Satisfaction and Constraint Logic Programming. X, 217 pages. 2005. (Subseries LNAI).

Vol. 3418: U. Brandes, T. Erlebach (Eds.), Network Analysis. XII, 471 pages. 2005.

Vol. 3416: M. Böhlen, J. Gamper, W. Polasek, M.A. Wimmer (Eds.), E-Government: Towards Electronic Democracy. XIII, 311 pages. 2005. (Subseries LNAI).

Vol. 3415: P. Davidsson, B. Logan, K. Takadama (Eds.), Multi-Agent and Multi-Agent-Based Simulation. X, 265 pages. 2005. (Subseries LNAI).

Vol. 3414: M. Morari, L. Thiele (Eds.), Hybrid Systems: Computation and Control. XII, 684 pages. 2005.

Vol. 3412: X. Franch, D. Port (Eds.), COTS-Based Software Systems. XVI, 312 pages. 2005.

Vol. 3411: S.H. Myaeng, M. Zhou, K.-F. Wong, H.-J. Zhang (Eds.), Information Retrieval Technology. XIII, 337 pages. 2005.

Vol. 3410: C.A. Coello Coello, A. Hernández Aguirre, E. Zitzler (Eds.), Evolutionary Multi-Criterion Optimization. XVI, 912 pages. 2005.

Vol. 3409: N. Guelfi, G. Reggio, A. Romanovsky (Eds.), Scientific Engineering of Distributed Java Applications. X, 127 pages. 2005.

Vol. 3408: D.E. Losada, J.M. Fernández-Luna (Eds.), Advances in Information Retrieval. XVII, 572 pages. 2005.

Table of Contents

Regular Papers

Extracting Semantic Taxonomies of Nouns from a Korean MRD Using a Small Bootstrapping Thesaurus and a Machine Learning Approach*

SeonHwa Choi and HyukRo Park

Dept. of Computer Science, Chonnam National University,
300 Youngbong-Dong, Puk-Ku Gwangju, 500-757, Korea
csh123@dreamwiz.com, hyukro@chonnam.ac.kr

Abstract. Most approaches for extracting hypernyms of a noun from the definition in an MRD rely on the lexico-syntactic patterns compiled by human experts. Not only these methods require high cost for compiling lexico-syntatic patterns but also it is very difficult for human experts to compile a set of lexical-syntactic patterns with a broad-coverage, because in natural languages there are various different expressions which represent the same concept. To alleviate these problems, this paper proposes a new method for extracting hypernyms of a noun from an MRD. In proposed approach, we use only syntactic(part-of-speech) patterns instead of lexico-syntactic patterns in identifying hypernyms to reduce the number of patterns while keeping their coverage broad. Our experiment shows that the classification accuracy of the proposed method is 92.37% which is significantly much better than those of previous approaches.

1 Introduction

The importance of broad-coverage lexical/semantic knowledge-bases has been stressed more than ever before as the natural language processing (NLP) systems became large and applied to wide variety of application domains. These lexical/semantic knowledge-bases include such as lexicons, thesauri and ontologies and machine-readable dictionaries. A lexical/semantic knowledge-base contains a list of terms, their semantic definitions and some of the relationships that exist between terms such as synonym, antonym and hypernym. Among the various relationships between terms, many researchers believe that the taxonomy relationship is especially useful because it can be utilized in various inference processes found in machine translation, information retrieval, word sense disambiguation and so on.

The taxonomy relationship is usually represented in thesauri as the broad-term (BT) narrow-term (NT) relations. However, because building broad-coverage thesauri is a very costly and time-consuming job, they are not readily available and often too general to be applied to a specific domain.

The work presented here is an attempt to alleviate this problem by devising a method for constructing taxonomy relations automatically from a machine readable dictionary (MRD). We use semantic hierarchies of nouns in a small thesaurus and a definition of a noun in a Korean MRD.

* This study was financially supported by special research fund of Chonnam National University in 2004.

A. Montoyo et al. (Eds.): NLDB 2005, LNCS 3513, pp. 1–9, 2005.

Most of the previous approaches for extracting hypernyms of a noun from the definition in an MRD rely on the lexico-syntactic patterns compiled by human experts. Not only these methods require high cost for compiling lexico-syntactic patterns but also it is very difficult for human experts to compile a set of lexical-syntactic patterns with a broad-coverage because, in natural languages, there are various different expressions which represent the same concept. Accordingly the applicable scope of a set of lexico-syntactic patterns compiled by human is very limited.

To overcome the drawbacks of human-compiled lexico-syntactic patterns, we use part-of-speech (POS) patterns only and try to induce these patterns automatically using a small bootstrapping thesaurus and machine learning methods.

The rest of the paper is organized as follows. We introduce the related works to in section 2. Section 3 deals with the problem of features selection. In section 4, our problem is formally defined as a machine learning method and discuss implementation details. Section 5 is devoted to experimenal result. Finally, we come to the conclusion of this paper in section 6.

2 Related Work

[3] introduced a method for the automatic acquisition of the hyponymy lexical relation from unrestricted text, and gives several examples of lexico-syntactic patterns for hyponymy that can be used to detect these relationships including those used here, along with an algorithm for identifying new patterns. Her approach is complementary to statistically based approaches that find semantic relations between terms, in that hers requires a single specially expressed instance of a relation while the others require a statistically significant number of generally expressed relations. The hyponym-hypernym pairs found by Hearst's algorithm include some that she describes as "context and point-of-view dependent", such as "Washington/nationalist" and "aircraft/target". [4] was somewhat less sensitive to this kind of problem since only the most common hypernym of an entire cluster of nouns is reported, so much of the noise is filtered. [3] tried to discover new patterns for hyponymy by hand, nevertheless it is a costly and time-consuming job. In the case of [3] and [4], since the hierarchy is learned from text, it get to be domain-specific different from a general-purpose resource such as WordNet.

[2] proposed a method that combines a set of unsupervised algorithms in order to accurately build large taxonomies from any MRD, and a system that 1)performs fully automatic extraction of taxonomic links from MRD entries and 2) ranks the extracted relations in a way that selective manual refinement is allowed. In this project, he introduced the idea of the hyponym-hypernym relationship appears between the entry word and the genus term. Thus, usually a dictionary definition is written to employ a genus term combined with differentia which distinguishes the word being defined from other words with the same genus term. He finds the genus term by simple heuristic defined using several examples of lexico-syntactic patterns for hyponymy.

[1] presented the method to extract semantic information from standard dictionary definitions. His automated mechanism for finding the genus terms is based on the observation that the genus term from verb and noun definitions is typically the head of the defining phrase. The syntax of the verb phrase used in verb definitions makes it possible to locate its head with a simple heuristic: the head is the single verb follow-

ing the word *to*. He asserted that heads are bounded on the left and right by specific lexical defined by human intuition, and the substring after eliminating boundary words from definitions is regarded as a head.

By the similar idea to [2], [10] introduced six kinds of rule extracting a hypernym from Korean MRD according to a structure of a dictionary definition. In this work, she proposed that only a subset of the possible instances of the hypernym relation will appear in a particular form, and she divides a definition sentence into a head term combined with differentia and a functional term. For extracting a hypernym, she analyzes a definition of a noun by word list and the position of words, and then searches a pattern coinciding with the lexico-syntactic patterns made by human intuition in the definition of any noun, and then extracts a hypernym using an appropriate rule among 6 rules. For example, rule 2 states that if a word X occurs in front of a lexical pattern "*leul bu-leu-deon i-leum* (*the name to call*)",then X is extracted as a hypernym of the entry word.

Several approaches[11][12][13] have been researched for building a semantic hierarchy of Korean nouns adopting the method of [2].

3 Features for Hypernym Identification

Machine learning approaches require an example to be represented as a feature vector. How an example is represented or what features are used to represent the example has profound impact on the performance of the machine learning algorithms. This section deals with the problems of feature selection with respect to characteristics of Korean for successful identification of hypernyms.

Location of a Word. In Korean, a head word usually appears after its modifying words. Therefore a head word has tendency to be located at the end of a sentence. In the definition sentences in a Korean MRD, this tendency becomes much stronger. In the training examples, we found that 11% of the hypernyms appeared at the start, 81% of them appeared at the end and 7% appeared at the middle of a definition sentence. Thus, the location of a noun in a definition sentences is an important feature for determining whether the word is a hypernym or not.

POS of a Function Word Attached to a Noun. Korean is an agglutinative language in which a word-phrase is generally a composition of a content word and some number of function words. A function word denotes the grammatical relationship between word-phrases, while a content word contains the central meaning of the word-phrase.

In the definition sentences, the function words which attached to hypernyms are confined to a small number of POSs. For example, nominalization endings, objective case postpositions come frequently after hypernyms but dative postpositions or locative postpositions never appear after hypernyms. A functional word is appropriate feature for identifying hypernyms.

Context of a Noun. The context in which a word appears is valuable information and a wide variety of applications such as word clustering or word sense disambiguation make use of it. Like in many other applications, context of a noun is important in deciding hyperhyms too because hypernyms mainly appear in some limited context.

Although lexico-syntactic patterns can represent more specific contexts, building set of lexco-syntactic patterns requires enormous training data. So we confined ourselves only to syntactic patterns in which hypernyms appear.

We limited the context of a noun to be 4 word-phrases appearing around the noun. Because the relations between word-phrases are represented by the function words of these word-phrases, the context of a noun includes only POSs of the function words of the its neighboring word-phrases. When a word-phrase has more than a functional morpheme, a representative functional morpheme is selected by an algorithm proposed by [8].

When a noun appears at the start or at the end of a sentence, it does not have right or left context respectively. In this case, two treatments are possible. The simplest approach is to treat the missing context as don't care terms. On the other hand, we could extend the range of available context to compensate the missing context. For example, the context of a noun at the start of a sentence includes 4 POSs of function words in its right-side neighboring word-phrases.

4 Learning Classification Rules

Decision tree learning is one of the most widely used and a practical methods for inductive inference such as ID3, ASSISTANT, and C4.5[14]. Because decision tree learning is a method for approximating discrete-valued functions that is robust to noisy data, it has therefore been applied to various classification problems successfully.

Our problem is to determine for each noun in definition sentences of a word whether it is a hypernym of the word or not. Thus our problem can be modeled as two-category classification problem. This observation leads us to use a decision tree learning algorithm C4.5.

Our learning problem can be formally defined as followings:

- Task T: determining whether a noun is a hypernym of an entry word or not .
- Performance measure P: percentage of nouns correctly classified.
- Training examples E: a set of nouns appearing in the definition sentences of the MRD with their feature vectors and target values.

To collect training examples, we used a Korean MRD provided by Korean Term-Bank Project[15] and a Korean thesaurus compiled by Electronic Communication Research Institute. The dictionary contains approximately 220,000 nouns with their definition sentences while the thesaurus has approximately 120,000 nouns and taxonomy relations between them. The fact that 46% of nouns in the dictionary are missing from the thesaurus illustrates the necessity of this research i.e. to extend a thesaurus using an MRD.

Using the thesaurus and the MRD, we found that 107,000 nouns in the thesaurus have their hypernyms in the definition sentences in the MRD. We used 70% of these nouns as training data and the remaining 30% of them as evaluation data. For each training pair of hypernym/hyponym nouns, we build a triple in the form of (hyponym definition-sentences hypernym) as follows.

ga-gyeong	[a-leum-da-un gyeong-chi *(a beautiful scene)*]	gyeong-chi
hyponym	definition sentence	hypernym