Yasubumi Sakakibara
Satoshi Kobayashi
Kengo Sato
Tetsuro Nishino
Etsuji Tomita (Eds.)

LNAI 4201

Grammatical Inference: Algorithms and Applications

8th International Colloquium, ICGI 2006
Tokyo, Japan, September 2006
Proceedings

Springer

Yasubumi Sakakibara   Satoshi Kobayashi
Kengo Sato   Tetsuro Nishino
Etsuji Tomita (Eds.)

Grammatical Inference: Algorithms and Applications

8th International Colloquium, ICGI 2006
Tokyo, Japan, September 20-22, 2006
Proceedings

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Yasubumi Sakakibara
Keio University, Yokohama, Japan
E-mail: yasu@bio.keio.ac.jp

Satoshi Kobayashi
University of Electro-Communications, Tokyo, Japan
E-mail: satoshi@cs.uec.ac.jp

Kengo Sato
National Institute of Advanced Industrial Science and Technology, Tokyo, Japan
E-mail: sato-kengo@aist.go.jp

Tetsuro Nishino
University of Electro-Communications, Tokyo, Japan
E-mail: nishino@ice.uec.ac.jp

Etsuji Tomita
University of Electro-Communications, Tokyo, Japan
E-mail: tomita@ice.uec.ac.jp

Library of Congress Control Number: 2006932583

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

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-45264-8 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-45264-5 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11872436 06/3142 5 4 3 2 1 0

Lecture Notes in Artificial Intelligence 4201

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Lecture Notes in Artificial Intelligence (LNAI)

Vol. 4213: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), Knowledge Discovery in Databases: PKDD 2006. XXII, 660 pages. 2006.

Vol. 4212: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), Machine Learning: ECML 2006. XXIII, 851 pages. 2006.

Vol. 4201: Y. Sakakibara, S. Kobayashi, K. Sato, T. Nishino, E. Tomita (Eds.), Grammatical Inference: Algorithms and Applications. XII, 359 pages. 2006.

Vol. 4196: K. Fischer, E. André, I.J. Timm, N. Zhong (Eds.), Multiagent System Technologies. X, 185 pages. 2006.

Vol. 4188: P. Sojka, I. Kopeček, K. Pala (Eds.), Text, Speech and Dialogue. XIV, 721 pages. 2006.

Vol. 4183: J. Euzenat, J. Domingue (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 291 pages. 2006.

Vol. 4180: M. Kohlhase, OMDoc – An Open Markup Format for Mathematical Documents [version 1.2]. XIX, 428 pages. 2006.

Vol. 4160: M. Fisher, W.v.d. Hoek, B. Konev, A. Lisitsa (Eds.), Logics in Artificial Intelligence. XII, 516 pages. 2006.

Vol. 4155: O. Stock, M. Schaerf (Eds.), Reasoning, Action and Interaction in AI Theories and Systems. XVIII, 343 pages. 2006.

Vol. 4149: M. Klusch, M. Rovatsos, T.R. Payne (Eds.), Cooperative Information Agents X. XII, 477 pages. 2006.

Vol. 4139: T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala, Advances in Natural Language Processing. XVI, 771 pages. 2006.

Vol. 4133: J. Gratch, M. Young, R. Aylett, D. Ballin, P. Olivier (Eds.), Intelligent Virtual Agents. XIV, 472 pages. 2006.

Vol. 4130: U. Furbach, N. Shankar (Eds.), Automated Reasoning. XV, 680 pages. 2006.

Vol. 4120: J. Calmet, T. Ida, D. Wang (Eds.), Artificial Intelligence and Symbolic Computation. XIII, 269 pages. 2006.

Vol. 4114: D.-S. Huang, K. Li, G.W. Irwin (Eds.), Computational Intelligence, Part II. XXVII, 1337 pages. 2006.

Vol. 4108: J.M. Borwein, W.M. Farmer (Eds.), Mathematical Knowledge Management. VIII, 295 pages. 2006.

Vol. 4106: T.R. Roth-Berghofer, M.H. Göker, H. A. Güvenir (Eds.), Advances in Case-Based Reasoning. XIV, 566 pages. 2006.

Vol. 4099: Q. Yang, G. Webb (Eds.), PRICAI 2006: Trends in Artificial Intelligence. XXVIII, 1263 pages. 2006.

Vol. 4095: S. Nolfi, G. Baldassarre, R. Calabretta, J.C. T. Hallam, D. Marocco, J.-A. Meyer, O. Miglino, D. Parisi (Eds.), From Animals to Animats 9. XV, 869 pages. 2006.

Vol. 4093: X. Li, O.R. Zaïane, Z. Li (Eds.), Advanced Data Mining and Applications. XXI, 1110 pages. 2006.

Vol. 4092: J. Lang, F. Lin, J. Wang (Eds.), Knowledge Science, Engineering and Management. XV, 664 pages. 2006.

Vol. 4088: Z.-Z. Shi, R. Sadananda (Eds.), Agent Computing and Multi-Agent Systems. XVII, 827 pages. 2006.

Vol. 4087: F. Schwenker, S. Marinai (Eds.), Artificial Neural Networks in Pattern Recognition. IX, 299 pages. 2006.

Vol. 4068: H. Schärfe, P. Hitzler, P. Øhrstrøm (Eds.), Conceptual Structures: Inspiration and Application. XI, 455 pages. 2006.

Vol. 4065: P. Perner (Ed.), Advances in Data Mining. XI, 592 pages. 2006.

Vol. 4062: G. Wang, J.F. Peters, A. Skowron, Y. Yao (Eds.), Rough Sets and Knowledge Technology. XX, 810 pages. 2006.

Vol. 4049: S. Parsons, N. Maudet, P. Moraitis, I. Rahwan (Eds.), Argumentation in Multi-Agent Systems. XIV, 313 pages. 2006.

Vol. 4048: L. Goble, J.-J.C.. Meyer (Eds.), Deontic Logic and Artificial Normative Systems. X, 273 pages. 2006.

Vol. 4045: D. Barker-Plummer, R. Cox, N. Swoboda (Eds.), Diagrammatic Representation and Inference. XII, 301 pages. 2006.

Vol. 4031: M. Ali, R. Dapoigny (Eds.), Advances in Applied Artificial Intelligence. XXIII, 1353 pages. 2006.

Vol. 4029: L. Rutkowski, R. Tadeusiewicz, L.A. Zadeh, J.M. Zurada (Eds.), Artificial Intelligence and Soft Computing – ICAISC 2006. XXI, 1235 pages. 2006.

Vol. 4027: H.L. Larsen, G. Pasi, D. Ortiz-Arroyo, T. Andreasen, H. Christiansen (Eds.), Flexible Query Answering Systems. XVIII, 714 pages. 2006.

Vol. 4021: E. André, L. Dybkjær, W. Minker, H. Neumann, M. Weber (Eds.), Perception and Interactive Technologies. XI, 217 pages. 2006.

Vol. 4020: A. Bredenfeld, A. Jacoff, I. Noda, Y. Takahashi (Eds.), RoboCup 2005: Robot Soccer World Cup IX. XVII, 727 pages. 2006.

Vol. 4013: L. Lamontagne, M. Marchand (Eds.), Advances in Artificial Intelligence. XIII, 564 pages. 2006.

Vol. 4012: T. Washio, A. Sakurai, K. Nakajima, H. Takeda, S. Tojo, M. Yokoo (Eds.), New Frontiers in Artificial Intelligence. XIII, 484 pages. 2006.

Vol. 4008: J.C. Augusto, C.D. Nugent (Eds.), Designing Smart Homes. XI, 183 pages. 2006.

Vol. 4005: G. Lugosi, H.U. Simon (Eds.), Learning Theory. XI, 656 pages. 2006.

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

Vol. 3963: O. Dikenelli, M.-P. Gleizes, A. Ricci (Eds.), Engineering Societies in the Agents World VI. XII, 303 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.

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

Vol. 3949: F. A. Savacı (Ed.), Artificial Intelligence and Neural Networks. IX, 227 pages. 2006.

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

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

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

Vol. 3913: O. Boissier, J. Padget, V. Dignum, G. Lindemann, E. Matson, S. Ossowski, J.S. Sichman, J. Vázquez-Salceda (Eds.), Coordination, Organizations, Institutions, and Norms in Multi-Agent Systems. XII, 259 pages. 2006.

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

Vol. 3904: M. Baldoni, U. Endriss, A. Omicini, P. Torroni (Eds.), Declarative Agent Languages and Technologies III. XII, 245 pages. 2006.

Vol. 3900: F. Toni, P. Torroni (Eds.), Computational Logic in Multi-Agent Systems. XVII, 427 pages. 2006.

Vol. 3899: S. Frintrop, VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. XIV, 216 pages. 2006.

Vol. 3898: K. Tuyls, P.J. 't Hoen, K. Verbeeck, S. Sen (Eds.), Learning and Adaption in Multi-Agent Systems. X, 217 pages. 2006.

Vol. 3891: J.S. Sichman, L. Antunes (Eds.), Multi-Agent-Based Simulation VI. X, 191 pages. 2006.

Vol. 3890: S.G. Thompson, R. Ghanea-Hercock (Eds.), Defence Applications of Multi-Agent Systems. XII, 141 pages. 2006.

Vol. 3885: V. Torra, Y. Narukawa, A. Valls, J. Domingo-Ferrer (Eds.), Modeling Decisions for Artificial Intelligence. XII, 374 pages. 2006.

Vol. 3881: S. Gibet, N. Courty, J.-F. Kamp (Eds.), Gesture in Human-Computer Interaction and Simulation. XIII, 344 pages. 2006.

Vol. 3874: R. Missaoui, J. Schmidt (Eds.), Formal Concept Analysis. X, 309 pages. 2006.

Vol. 3873: L. Maicher, J. Park (Eds.), Charting the Topic Maps Research and Applications Landscape. VIII, 281 pages. 2006.

Vol. 3864: Y. Cai, J. Abascal (Eds.), Ambient Intelligence in Everyday Life. XII, 323 pages. 2006.

Vol. 3863: M. Kohlhase (Ed.), Mathematical Knowledge Management. XI, 405 pages. 2006.

Vol. 3862: R.H. Bordini, M. Dastani, J. Dix, A.E.F. Seghrouchni (Eds.), Programming Multi-Agent Systems. XIV, 267 pages. 2006.

Vol. 3849: I. Bloch, A. Petrosino, A.G.B. Tettamanzi (Eds.), Fuzzy Logic and Applications. XIV, 438 pages. 2006.

Vol. 3848: J.-F. Boulicaut, L. De Raedt, H. Mannila (Eds.), Constraint-Based Mining and Inductive Databases. X, 401 pages. 2006.

Vol. 3847: K.P. Jantke, A. Lunzer, N. Spyratos, Y. Tanaka (Eds.), Federation over the Web. X, 215 pages. 2006.

Vol. 3835: G. Sutcliffe, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XIV, 744 pages. 2005.

Vol. 3830: D. Weyns, H. V.D. Parunak, F. Michel (Eds.), Environments for Multi-Agent Systems II. VIII, 291 pages. 2006.

Vol. 3817: M. Faundez-Zanuy, L. Janer, A. Esposito, A. Satue-Villar, J. Roure, V. Espinosa-Duro (Eds.), Nonlinear Analyses and Algorithms for Speech Processing. XII, 380 pages. 2006.

Vol. 3814: M. Maybury, O. Stock, W. Wahlster (Eds.), Intelligent Technologies for Interactive Entertainment. XV, 342 pages. 2005.

Vol. 3809: S. Zhang, R. Jarvis (Eds.), AI 2005: Advances in Artificial Intelligence. XXVII, 1344 pages. 2005.

Vol. 3808: C. Bento, A. Cardoso, G. Dias (Eds.), Progress in Artificial Intelligence. XVIII, 704 pages. 2005.

Vol. 3802: Y. Hao, J. Liu, Y.-P. Wang, Y.-m. Cheung, H. Yin, L. Jiao, J. Ma, Y.-C. Jiao (Eds.), Computational Intelligence and Security, Part II. XLII, 1166 pages. 2005.

Vol. 3801: Y. Hao, J. Liu, Y.-P. Wang, Y.-m. Cheung, H. Yin, L. Jiao, J. Ma, Y.-C. Jiao (Eds.), Computational Intelligence and Security, Part I. XLI, 1122 pages. 2005.

Vol. 3789: A. Gelbukh, Á. de Albornoz, H. Terashima-Marín (Eds.), MICAI 2005: Advances in Artificial Intelligence. XXVI, 1198 pages. 2005.

Vol. 3782: K.-D. Althoff, A. Dengel, R. Bergmann, M. Nick, T.R. Roth-Berghofer (Eds.), Professional Knowledge Management. XXIII, 739 pages. 2005.

Vol. 3763: H. Hong, D. Wang (Eds.), Automated Deduction in Geometry. X, 213 pages. 2006.

Preface

The 8th International Colloquium on Grammatical Inference (ICGI 2006) was held at the University of Electro-Communications (UEC), Tokyo, Japan on September 20-22, 2006. ICGI 2006 was the eighth in a series of successful biennial international conferences in the area of grammatical inference. Previous meetings were held in Essex, UK; Alicante, Spain; Montpellier, France; Ames, Iowa, USA; Lisbon, Portugal; Amsterdam, Netherlands; Athens, Greece. ICGI 2006 was the first conference in this series to be held in Asia. This series of conferences seeks to provide a forum for presentation and discussion of original research papers on all aspects of grammatical inference.

Grammatical inference, the study of learning grammars from data, is an established research field in artificial intelligence, dating back to the 1960s and has been extensively addressed by researchers in automata theory, language acquisition, computational linguistics, machine learning, pattern recognition, computational learning theory and neural networks. ICGI 2006 successively emphasized on the multi-disciplinary nature of the research field and the diverse domains in which grammatical inference is being applied, such as natural language acquisition, computational biology, structural pattern recognition, information retrieval, Web mining, text processing, data compression and adaptive intelligent agents.

We received 44 high-quality papers from 14 countries around the world. The papers were reviewed by three reviewers. Based on the positive comments of the reviewers, 25 full papers were accepted. In addition, we decided to accept 8 short papers for poster presentation. Short papers appear as two-page extended abstracts in a separate section of this volume. The topics of the accepted papers vary from theoretical results of learning algorithms to innovative applications of grammatical inference, and from learning several interesting classes of formal grammars to applications to natural language processing.

In parallel to the submission and reviewing of research papers, a machine translation competition, named Tenjinno, took place. In a separate paper in this volume, the organizers of the competition report on the peculiarities of such an endeavor and some interesting theoretical findings to which they have been led. Last but not least, we were honored by the contributions of our two invited speakers, Yuji Matsumoto, from Nara Institute of Science and Technology, Japan, and Jean-Philippe Vert, from Ecole des Mines de Paris, France. Both invited speakers provided interesting talks on the topics of natural language processing and bioinformatics, and we hope both talks invoked potential applications of grammatical inference.

The editors would like to acknowledge the contribution of the conference's Program Committee and the Additional Reviewers in reviewing the submitted papers and thank the Organizing Committee for their invaluable help in organizing the conference. Particularly, we would like to thank Colin de la Higuera, Menno van Zaannen, Bradford Starkie, and Dominique Estival for their additional voluntary service to the grammatical inference community, through this conference. We would also like to acknowledge the use of the Cyberchair software, from Borbala online conference services, in the submission and reviewing process. Finally, we are grateful for the generous support and sponsorship of the conference by the University of Electro-Communications, the PASCAL, Inoue foundation for Science, SIG Mathematical Modeling and Problem Solving in Information Processing Society of Japan and New Horizons in Computing (NHC) (Scientific Research on Priority Areas, supported by MEXT, Japan).

September 2006

Yasubumi Sakakibara
Satoshi Kobayashi
Kengo Sato
Tetsuro Nishino
Etsuji Tomita

Organization

Conference Chair

Etsuji Tomita	University of Electro-Communications, Japan

Technical Program Committee Co-chairs

Yasubumi Sakakibara	Keio University, Japan
Satoshi Kobayashi	University of Electro-Communications, Japan

Technical Program Committee

Naoki Abe	IBM Thomas J. Watson Research Center, USA
Pieter Adriaans	Perot Systems Corporation/University of Amsterdam, Netherlands
Dana Angluin	Yale University, USA
Hiroki Arimura	Hokkaido University, Japan
Mitra Basu	City University of New York, USA
François Coste	Symbiose, INRIA/IRISA, France
Pierre Dupont	University of Louvain, Belgium
Henning Fernau	University of Hertfordshire, UK
Colin de la Higuera	EURISE, Univ. de St. Etienne, France
Vasant Honavar	Iowa State University, USA
Chih-Jen Lin	National Taiwan University, Taiwan
Laurent Miclet	ENSSAT, Lannion, France
Gopalakrishnaswamy Nagaraja	Indian Institute of Technology, India
Katsuhiko Nakamura	Tokyo Denki University, Japan
Jacques Nicolas	IRISA, France
Tim Oates	University of Maryland Baltimore County, USA
Arlindo Oliveira	Lisbon Technical University, Portugal
Jose Oncina Carratala	Universidad de Alicante, Spain
Georgios Paliouras	Inst. of Informatics and Telecommunications, NCSR , Greece
Rajesh Parekh	Yahoo!, USA
Kengo Sato	CBRC, NAIST, Japan
Giora Slutzki	Iowa State University, USA
Bradford Starkie	Starkie Enterprise, Australia
Eiji Takimoto	Tohoku University, Japan
Menno van Zaanen	Universiteit van Amsterdam, Netherlands
Enrique Vidal	Universidad Politecnica de Valencia, Spain
Osamu Watanabe	Tokyo Institute of Technology, Japan
Thomas Zeugmann	Hokkaido University, Japan

Additional Reviewers

T. Armstrong
L. Becerra-Bonache
M. Bugalho
D. Eisenstat
J.-C. Janodet
H.-U. Krieger
J. A. Laxminarayana
A. Martins
J. Poland
J. M. Vilar

Organizing Committee Chair

Tetsuro Nishino	University of Electro-Communications, Japan

Organizing Committee

Colin de la Higuera	EURISE, Univ. de St. Etienne, France
Kazuhiro Hotta	University of Electro-Communications, Japan
Satoshi Kobayashi	University of Electro-Communications, Japan
Yoichi Motomura	National Institute of Advanced Industrial Science and Technology, Japan
Katsuhiko Nakamura	Tokyo Denki University, Japan
Seiya Okubo	University of Electro-Communications, Japan
Yasuhiro Tajima	Tokyo University of Agriculture and Technology, Japan
Haruhisa Takahashi	University of Electro-Communications, Japan
Jun Tarui	University of Electro-Communications, Japan
Mitsuo Wakatsuki	University of Electro-Communications, Japan

Sponsoring Institutions

University of Electro-Communications

PASCAL Network

Inoue Foundation for Science

SIG Mathematical Modeling and Problem Solving in Information Processing Society of Japan

New Horizons in Computing (NHC) (Scientific Research on Priority Areas, supported by MEXT, Japan)

Table of Contents

Invited Papers

Regular Papers

Poster Papers

Parsing Without Grammar Rules

Yuji Matsumoto

Graduate School of Information Science
Nara Institute of Science and Technology
Takayama, Ikoma, Nara 630-0192 Japan
matsu@is.naist.jp

Abstract. In this article, we present and contrast recent statistical approaches to word dependency parsing and lexicalized formalisms for grammar and semantics. We then consider the possibility of integrating those two extreme ideas, which leads to fully lexicalized parsing without any syntactic grammar rules.

Keywords: dependency parsing, lexicalized grammar, lexical semantics.

1 Introduction

Traditional syntactic analysis of natural languages mainly assumes a set of phrase structure grammar rules possibly with some syntactic information in the lexicon such as case frames. Then, parsing is done with phrase structure parsing algorithms such as Chart Parsing or CKY Parsing algorithms. In contrast, recent grammar formalisms, such as HPSG (Head-driven Phrase Structure Grammar)[15] and LTAG (Lexicalized Tree Adjoining Grammar)[14], originating from phrase structure-style grammars, are extremely lexicalized (termed radical lexicalism), and now have only a few grammar rules (or grammar schemes). In such systems, most of syntactic information is stored in lexical entries. On the other hand, lexical semantic theories such as LCS (Lexical Conceptual Structure)[6] and GL (Generative Lexicon)[13] propose to assume very rich semantic information in lexical entries in a language, and give a systematic explanation of syntactic ambiguities or syntactic alternation that are dealt with in traditional phrase structure-based analysis by describing multiple frames corresponding to each of possilbe syntactic constructions. Furthermore, constructions that are not assumed in the case frames of a word may appear in real language use. Construction grammar approaches to language[5][12] aim to explain such phenomena.

Recent trend of natural language parsing moved to corpus-based research, where a large scale parsed corpus is used to estimate statistical properties of language constructions. Early research in this direction[1][3] used to use phrase structure trees in their analysis as they base their syntactic structure on Penn Treebank[8]. More recently, word dependency parsing is getting larger attention [9][11][16] because of its simplicity and easiness in adaptability to various languages (e.g., this year's CoNLL shared task was multi-lingual dependency parsing[1]).

[1] `http://nextens.uvt.nl/~conll/`

Y. Sakakibara et al. (Eds.): ICGI 2006, LNAI 4201, pp. 1–6, 2006.

In this paper, we introduce those recent trends in lexicalism in both parsing domain and grammar formalisms, and discuss possible integration of these two extreme ideas.

2 Word Dependency Parsing

There is a traditional syntactic analysis for Japanese sentences, named bunsetsu dependency analysis. A bunsetsu means a base Japanese phrase consisting of content words followed by functional words/functional inflection form. The syntactic structure of a Japanese sentence can be represented by dependency relation between bunsetsu's. Only the conditions of this dependency are quite simple that dependency trees must be connected, single headed, acyclic and projective (no-crossing). An interesting characteristics of Japanese dependency structure is that any bunsetsu (except the right most one) modifies one of the bunsetsu's on its right side because Japanese is a head-final language. This makes it easy to construct a bunsetsu dependency parsing in a very simple way, and we proposed a Japanese deterministic dependency parser[7] based on Support Vector Machines.

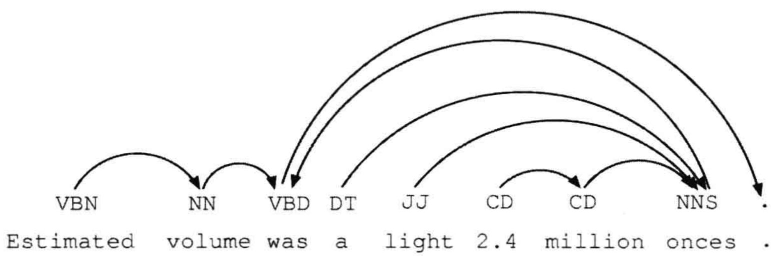

Fig. 1. An example of English projective dependency tree

We then extended this idea into a Shift-Reduce style depterministic parsing, and applied it to English and Chinese parsing[2][16]. Fig. 1 shows an example of English word dependency tree. The examples shown in this paper are unlabeled trees, while some dependency trees assume their edges to be labeled with something like SUBJ, OBJ, etc. In our approach, dependency relationship (left-direction, right-direction, or none) between two adjacent nodes (words) is deterministically decided as a classification task learned by Support Vector Machines, and the parsing is done from bottom to top. Because of robust learning ability of SVMs, the current parser can achieve more than 90% accuracy in practical English sentence analysis (for sentences in Penn Treebank). Nivre et al[11] took a similar approach to dependency tree analysis.

This year's CoNLL (Conference on Natural Language Learning) shared task was Multi-lingual Dependency Parsing, and the target was to built a corpus-based language independent dependecy parser and to test it with thirteen languages provided by the conference organizer. In many of languages there are non-projective sentences, in which some dependency relations cross each other (Fig. 2 shows one of such examples). This kind of sentences cannot be formulated by phrase structure grammars and are difficult to parse with the parsing algorithms originally designed for phrase structure grammars.

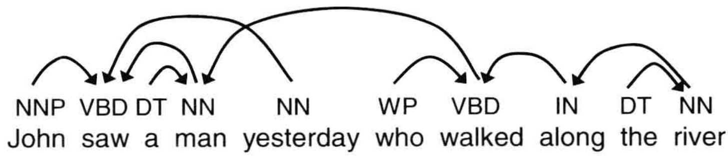

Fig. 2. An example of English non-projective dependency tree

Recent McDonald et al's work[10] showed that non-projected dependency analysis is easily formulated as a search problem for the maximum cost spanning tree.

Since resolution of syntactic ambiguity has been the most difficult problem in parsing natural language sentences, the advantage of those corpus-based or statistical approaches is its ability of disambiguation, that is, they produce the most plausible parse tree considering all the dependency relation appeared in the training corpus.

3 Lexicalized Grammars and Lexical Semantic Thoeries

3.1 Lexicalized Grammars and Dependency Parsing

As we explained in Introduction, recent lexicalized grammar formalisms put most of the grammatical information to the lexicon. In HPSG, each predicate (a verb, an adjective or an auxiliary verb) has argument structure that describes information of its complements. There are only a few grammar schemes such as head complement rule, head adjunct rule, head specifier rule, and so on, all of which can be specified as a binary tree where either one of them plays a role of syntactic head. In LTAG, every lexical entry is associated with a tree that shows its syntactic property. There are only two grammar rules or attachment rules, substitution and adjoining. The application of these rules can be depicted as a derivation tree, where one tree structure is attached to another with either of two attachment rules. In both of HPSG and LTAG, basic operations can be defined as a binary construction of a tree, which seems to have close relationship with word dependency structure. Although there may be some discrepancy between the binary relations in different systems, dependency parsing will give a good control information in syntactic parsing based on lexicalized grammar formalisms.

3.2 Lexical Semantics and Dependency Parsing

While lexical semantics theories such as Lexical Conceptual Structure[6] and Generative Lexicon[13] do not specify syntactic structure of a language, they