ACL-05

Tutorials

Abstracts

25 June 2005
University of Michigan
Ann Arbor, Michigan, USA

Production and Manufacturing by
Omnipress Inc.
Post Office Box 7214
Madison, WI 53707-7214

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
75 Paterson Street, Suite 9
New Brunswick, NJ 08901
USA
Tel: +1-732-342-9100
Fax: +1-732-342-9339
acl@aclweb.org

Tutorial Chair

Stefan Riezler, Palo Alto Research Center

Tutorials and Presenters

Morning

Advances in Word Sense Disambiguation
Rada Mihalcea and Ted Pedersen

Arabic Natural Language Processing
Nizar Habash

Empirical Methods for Dialogue System Research
Gregory Aist

Afternoon

Recent Developments in Computational Semantics
Valia Kordoni and Markus Egg

SVM's and Structured Max-Margin Methods
Dan Klein and Ben Taskar

ADVANCES IN WORD SENSE DISAMBIGUATION

Ted Pedersen and Rada Mihalcea
http://www.cs.unt.edu/~rada
http://www.d.umn.edu/~tpederse

Word Sense Disambiguation is the problem of identifying the intended meaning (or sense) of a word, based on the context in which it occurs. This is a central problem in natural language processing, and improved approaches have the potential to advance the state of the art in machine translation, information retrieval, and many other language related problems. This tutorial will introduce the full range of techniques that have been applied to this problem. These include knowledge-intensive methods that take advantage of dictionaries and other manually crafted resources, supervised techniques that learn classifiers from training examples, minimally supervised approaches that bootstrap off small amounts of labeled data, and unsupervised approaches that identify word senses in raw unannotated text. In addition, the tutorial will provide an overview of resources that are available to those who might wish to conduct research in this area, or incorporate word sense disambiguation techniques in their existing systems.

Tutorial attendees will come away with a firm understanding of all the major approaches to word sense disambiguation that are currently under investigation in the Computational Linguistics community. Our objective is that attendees will have sufficient understanding to make informed decisions about including word sense disambiguation techniques in their text processing applications in the future, and to see where there might be opportunities to advance the state of the art in word sense disambiguation by the application of novel techniques from their own areas of expertise.

This tutorial is intended for NLP researchers and practitioners who seek a general understanding of Word Sense Disambiguation. It is introductory in nature, no special knowledge or background is required.

Tutorial Outline

1. Introduction (Pedersen)
   (a) Word Sense Disambiguation Defined and Illustrated
   (b) Historical Overview
   (c) Practical Applications
2. Methodology (Mihalcea)
   (a) All Words Disambiguation
   (b) Targeted Words (Lexical Sample) Disambiguation
   (c) Word Sense Discrimination and Sense Discovery
   (d) Evaluation (granularity and scoring)

Tutorial Abstracts for the 43rd Annual Meeting of the ACL (Companion Volume), pages 1–3,
Ann Arbor, June 2005.

3. Knowledge Intensive Methods (Mihalcea)
    (a) Machine Readable Dictionaries
    (b) Selectional Restrictions
    (c) Measures of Semantic Similarity
    (d) Heuristic-based Methods
4. Supervised Learning Methods (Pedersen)
    (a) Introduction to Classifier Induction
    (b) Support Vector Machines in WSD
    (c) Ensemble Methods
5. Minimally Supervised Methods (Mihalcea)
    (a) Introduction to Bootstrapping and Co-Training
    (b) Yarowsky's Algorithm
    (c) Using the Web
6. Unsupervised Methods (Pedersen)
    (a) Discrimination as First Step of Disambiguation
    (b) Clustering Senses from Unannotated Corpora
    (c) Sense Discrimination Using Parallel Texts
7. How to Get Started in WSD Research (Mihalcea)
    (a) Software
    (b) Lexicons and Thesauruses
    (c) Sense Tagged Text
    (d) Senseval exercises
8. Conclusions (Pedersen)
    (a) The Web and WSD
    (b) Multilingual WSD
    (c) The Next Five Years

RADA MIHALCEA is an Assistant Professor of Computer Science at the University of North Texas. Her research interests are in lexical semantics, minimally supervised natural language learning, and graph-based algorithms for text processing. She is the president of the ACL Special Group on the Lexicon (SIGLEX) and a board member for the ACL Special Group on Natural Language Learning (SIGNLL). She was one of the coordinators of the Senseval-3 word sense disambiguation evaluation exercise.

TED PEDERSEN is an Associate Professor of Computer Science at the University of Minnesota, Duluth. He has been actively engaged in word sense disambiguation research since 1995. His work includes supervised machine learning approaches to word sense disambiguation, unsupervised clustering approaches to word sense discrimination, and disambiguation via the use of measures of semantic similarity and relatedness. He is the recipient of an NSF Faculty Early Development (CAREER) Award.

ARABIC NATURAL LANGUAGE PROCESSING

Nizar Habash
habash@cs.columbia.edu
http://www.cs.columbia.edu/~habash

This tutorial provides NLP system developers/researchers with necessary background information for working with the Arabic language, which has recently become a focus of an increasing number of projects in computational linguistics. The goal of the tutorial is to introduce Arabic linguistic phenomena that need to be addressed and review the state-of-the-art on Arabic processing. Alternative approaches are presented and contrasted for their value in different application contexts (e.g., information retrieval versus machine translation).

The tutorial has four sections. First is a discussion of Arabic phonology and orthography with a focus on Arabic spelling peculiarities and their effect on Arabic processing. Arabic encoding issues are also addressed. Second, aspects of Arabic morphology are presented and explained. This is followed by a survey of different approaches to address these phenomena. Third, a survey of Arabic syntactic phenomena is presented and contrasted to English syntactic phenomena. Syntactic representation in the Penn Arabic Treebank is discussed. Finally, Arabic dialects and the kind of problems they present for Arabic NLP are presented. Links to recent publications and available toolkits/ resources for all four sections are provided.

This tutorial is designed for computer scientists and linguistics alike. Acquaintance with basic formal language theory and knowledge of some programming language will be useful, but not mandatory.

Tutorial Outline

1. Arabic Orthography
   (a) Phonology
   (b) Orthography
   (c) Encoding Issues
2. Arabic Morphology
   (a) Introduction to Arabic Morphology
   (b) Arabic Morphological Analysis/Generation
3. Arabic Syntax
   (a) Arabic Syntactic Phenomena
   (b) Arabic Parsing Issues
4. Arabic Dialects
   (a) Introduction to Arabic Dialects

Tutorial Abstracts for the 43rd Annual Meeting of the ACL (Companion Volume), pages 4–5,
Ann Arbor, June 2005.

(b) Processing of Arabic dialects

NIZAR HABASH received his PhD in 2003 from the Computer Science Department, University of Maryland College Park. He is currently a researcher at the Center for Computational Learning Systems in Columbia University. His research includes work on machine translation, natural language generation, lexical semantics, and morphological analysis, generation and disambiguation. His work on Arabic ranges from research in Arabic encoding issues to Arabic-English machine translation and includes computational modeling of Arabic dialects for machine translation and speech recognition, and Arabic dialect parsing.

EMPIRICAL METHODS FOR DIALOGUE SYSTEM RESEARCH

Gregory Aist
http://www.gregoryaist.com/

This tutorial will present a comprehensive overview of empirical methods for building dialogue systems. We will cover phases of development including requirements gathering, design, implementation, testing, refinement, deployment, evaluation, customer support, and product maintenance. The emphasis throughout will be on how to turn design issues into empirical questions, and on learning a variety of techniques for answering those questions from data collected in the lab and in the field. We will spend about one-third of the time covering fundamental techniques, and the remainder on more advanced methods.

This tutorial will be timely for researchers with a background in a core area of NLP such as parsing who are planning to move into dialogue systems and wish to employ the same empirical rigor to dialogue as has become the norm in the parsing community. Those working in dialogue systems will benefit from an up-to-date overview of empirical methods across the wide range of development phases, particularly since methods for some phases such as testing, refinement, and support are less commonly known in the community. Students in NLP and in fields such as statistics or experimental design will be able to get a broad picture of the state of the art in applying empirical methods to dialogue systems. Finally, companies that are looking to launch a dialogue system effort will especially benefit from the requirements gathering and design sessions since we will present data-driven ways to figure out where in a business or a customer's interactions would substantially benefit from a dialogue systems application.

Tutorial Outline

1. First session
   (a) Introduction
   (b) Goals and Requirements
   (c) Design and Implementation
   (d) Questions and Discussion
2. Second Session
   (a) Testing and Refinement
   (b) Deployment and Evaluation
   (c) Support and Maintenance
   (d) Questions and Discussion
   (e) Concluding Remarks

GREGORY AIST is a Research Associate at the University of Rochester. His scientific interests are in language, learning and computation. Language: human lexical learning, multi-topic conversation.

Tutorial Abstracts for the 43rd Annual Meeting of the ACL (Companion Volume), pages 6–7,
Ann Arbor, June 2005.

Learning: effective strategies for instruction, language learning, the role of emotions in learning, learning procedural tasks. Computation: spoken language understanding, turn-taking, multimodal generation, intelligent tutoring systems, incrementality in spoken dialogue systems.

RECENT DEVELOPMENTS IN COMPUTATIONAL SEMANTICS

Markus Egg and Valia Kordoni
Rijksuniversiteit Groningen/Universität des Saarlandes
egg@let.rug.nl/kordoni@coli.uni-sb.de

The last decades have seen considerable progress and increasing interest in Computational Semantics (CS). Several semantic formalisms have been successfully implemented and integrated in large-scale NLP systems. At the same time, there is a increasing tendency for different approaches to CS to converge.

We use a suitable semantic formalism to introduce the central issues of CS and to outline the ideas and insights behind the implementations. After discussing the derivation of expressions of such formalisms by appropriate syntax-semantics interfaces we focus on the integration of CS formalisms in real-life applications and their evaluation.

Tutorial outline

1. introduction to the field of Computational Semantics
   - 1.1 motivation of CS systems: efficient handling of ambiguity by underspecified representations
   - 1.2 central issues of CS (in particular, scope, ellipsis, anaphora and their interaction)
2. syntax-semantics interface
   - 2.1 efficient derivation of semantic representations
   - 2.2 portability of CS modules to different systems
3. CS formalisms in real life
   - 3.1 CS components of large-scale, large-coverage grammars
   - 3.2 CS for QA systems, Information Retrieval, Information Management, and Hyperlinking
   - 3.3 CS for hybrid (deep and shallow) processing
   - 3.4 evaluation of and translatability between CS systems

MARKUS EGG works as a Reader at the University of Groningen. He is interested in syntax, semantics, and pragmatics, with an emphasis on the question of how meaning is first constructed on the basis of surface-oriented, tractable syntax formalisms, and then augmented to a fully-fledged utterance meaning through the application of pragmatic and encyclopaedic knowledge. He works on the theoretical development of these areas as well as on their implementation in NLP systems.

VALIA KORDONI is a Senior Researcher at Saarland University and the LT Lab of DFKI GmbH. She is interested in syntax, semantics and the syntax-semantics interface. She has worked on the theoretical development of the aforementioned, as well as on their implementation in NLP systems.

Tutorial Abstracts for the 43rd Annual Meeting of the ACL (Companion Volume), page 8,
Ann Arbor, June 2005.

SVMS AND STRUCTURED MAX-MARGIN METHODS

Dan Klein and Ben Taskar
http://www.cs.berkeley.edu/~klein
http://www.cs.berkeley.edu/~taskar

This tutorial will be divided into two parts. Part I will be an introduction to SVMs from the ground up, emphasizing how max-margin classifiers relate to maximum entropy classifiers. It will present the arguments for and against max-margin methods in several ways, describe duality in a self-contained way, and discuss what kernels are and aren't. Multiclass SVM formulations and their associated optimization methods will be presented. The goal of this part of the tutorial is to increase awareness of the strengths, weaknesses, and inner workings of SVMs as classifiers for NLP tasks. This part will be most useful (1) to researchers to whom SVMs are currently a magic black box and (2) as a point of departure for part II. Part II will describe structured max-margin methods. The presentation will first be to work out structured solutions in the framework of part I and show how this solution would work (if infeasibly slowly). Then, the max-margin dynamic programming analogs to standard dynamic programs used for calculating expectations will be presented, along with the structured optimization methods. Kernelization and max-margin approaches to combinatorial problems other than sequences and trees will be mentioned, though in less detail. The tutorial will conclude with a discussion of some max-margin learning software we will be making available.

Tutorial Outline

1. SVMs
   (a) Overview: discriminative feature-based classifiers in NLP
   (b) Different formulations lead to different classifiers: comparison of SVMs and MaxEnt models
   (c) Multiclass SVMs: ranking constraints and the associated optimization problem
   (d) Duality: introduction to duality, dual classifiers, why duality is useful
   (e) Optimization methods in medium detail, focus on SMO / coordinate ascent
   (f) Experiments with comparative data analysis
   (g) Introduction to kernels (but not an extensive coverage of known kernels)
2. Structured Max-Margin Methods
   (a) Structured loss functions
   (b) Exhaustive approach to structured problems
   (c) Dynamic programming and max-marginal calculations
   (d) Optimization in structured models
   (e) Example experiments and results: sequence and tree models

Tutorial Abstracts for the 43rd Annual Meeting of the ACL (Companion Volume), pages 9–10,
Ann Arbor, June 2005.

(f) Compare / contrast with HMMs, CRFs, PCFGs

(g) Kernelization

(h) Other max-margin problems: matchings, tours

(i) Overview of software resources

DAN KLEIN is an assistant professor in the Computer Science Division at UC Berkeley. He received his bachelors degree (summa cum laude) from Cornell University. He then went to Oxford University on a Marshall scholarship, where he earned a masters degree in linguistics, and finally to Stanford University for his masters and PhD in computer science. Prof. Klein's research interests include efficient machine learning algorithms for complex NLP tasks, unsupervised learning of linguistic structure, and integrating linguistic ideas into machine learning approaches to NLP. He was the recipent of best paper awards at ACL 2003 and EMNLP 2004.

BEN TASKAR went on to a postdoctoral fellowship at the Computer Science Division, University of California at Berkeley, after receiving his Ph.D. in Computer Science from Stanford University. One of his primary interests is structured models in machine learning, especially in computational linguistics, computer vision and computational biology. Last year, he has co-organized a NIPS workshop on this topic. His work on structured prediction in sequence, tree and other models has received best paper awards at NIPS and EMNLP conferences.