HLT-NAACL 2006

Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics

Tutorial Abstracts

Chris Manning, Doug Oard and Jim Glass
Tutorial Chairs

June 4-9, 2006
New York City, USA

Published by the Association for Computational Linguistics
http://www.aclweb.org

Table of Contents

1. What's in a Name: Current Methods, Applications, and Evaluation in Multilingual Name Search and Matching

Sherri Condon and Keith J. Miller, MITRE

Names of people, places, and organizations have unique linguistic properties, and they typically require special treatment in automatic processes. Appropriate processing of names is essential to achieve high-quality information extraction, speech recognition, machine translation, and information management, yet most HLT applications provide limited specialized processing of names. Variation in the forms of names can make it difficult to retrieve names from data sources, to perform co-reference resolution across documents, or to associate instances of names with their representations in gazetteers and lexicons. Name matching has become critical in government contexts for checking watchlists and maintaining tax, health, and Social Security records. In commercial contexts, name matching is essential in credit, insurance, and legal applications.

This tutorial will focus on personal names, with special attention given to Arabic names, though it will be clear that much of the material applies to other languages and to names of places and organizations. Case studies will be used to illustrate problems and approaches to solutions. Arabic names illustrate many of the issues encountered in multilingual name matching, among which are complex name structures and spelling variation due to morphophonemic alternation and competing transliteration conventions.

1.1 Tutorial Outline

1. Name matching across languages, scripts, and cultures
   - Survey of problems using Arabic case study
     * Name parts and structure (titles, initials, particles, prefixes, suffixes, nicknames, tribal names)
     * Transliteration complications (segmentation, ambiguity, incompleteness, dialect variation, acoustic mismatches, competing standards)
     * Other difficulties presented by personal names
   - Survey of approaches to solutions, advantages/disadvantages of each:
     * SOUNDEX, generic string matching (Levenshtein, n-gram, Jaro-Winkler),
     * Variant generation (pattern matching, dictionaries, gazetteers),
     * Normalization (morphological analysis, rewriting, "deep" structures)
     * Intelligent-search algorithms that incorporate linguistic knowledge in selection of string-similarity measures, parameters, and lists
   - Matching across scripts
     * Methods for data acquisition
     * Transliteration
     * Phonological interlingua
2. Evaluation of Name Search and Matching Systems
   - Development of ground-truth sets
     * Human adjudication
     * Estimation techniques
   - Case study: adjudication exercises
   - Issues in establishing ground truth: different truth for different applications
   - Metrics (precision, recall, F scores, others)
   - Case study comparing matching systems for Romanized Arabic names (based on MITRE evaluation of 9 name matching products)
   - Inter-adjudicator agreement
   - Performance and other considerations

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 299–300,
New York City, June 2006.

1.2 Target Audience

This tutorial is intended for those with interest in information retrieval and entity extraction, identity resolution, Arabic computational linguistics, and related language-processing applications. As a relatively unstudied domain, name matching is a promising area for innovation and for researchers seeking new projects.

Keith J. Miller received his Ph.D. in Computational Linguistics from Georgetown University. He spent several years working on various large-scale name matching systems. His current research activities center around multicultural name matching, machine translation, embedded HLT systems, and component and system-level evaluation of systems involving HLT components.

Sherri Condon received her Ph.D. in Linguistics from the University of Texas at Austin. In addition to several years of work in multilingual name matching and cross script name matching, she is a researcher in discourse/dialogue, entity extraction, and evaluation of machine translation and dialogue systems.

2. Beyond EM: Bayesian Techniques for Human Language Technology Researchers

Hal Daume III, USC-ISI

The Expectation-Maximization (EM) algorithm has proved to be a great and useful technique for unsupervised learning problems in natural language, but, unfortunately, its range of applications is largely limited by intractable E- or M-steps, and its reliance on the maximum likelihood estimator. The natural language processing community typically resorts to ad-hoc approximation methods to get (some reduced form of) EM to apply to our tasks. However, many of the problems that plague EM can be solved with Bayesian methods, which are theoretically more well justified. This tutorial will cover Bayesian methods as they can be used in natural language processing. The two primary foci of this tutorial are specifying prior distributions and performing the necessary computations to perform inference in Bayesian models. The focus of the tutorial will be primarily on unsupervised techniques (for which EM is the obvious choice). Supervised and discriminative techniques will also be mentioned at the conclusion of the tutorial, and pointers to relevant literature will be provided.

2.1 Tutorial Outline

1. Introduction to the Bayesian Paradigm
2. Background Material
   - Graphical Models (naive Bayes, maximum entropy, HMMs)
   - Expectation Maximization
   - Non-Bayesian Inference Techniques
3. Common Statistical Distributions
   - Uniform
   - Binomial and Multinomial
   - Beta and Dirichlet
   - Poisson, Gaussian and Gamma
4. Simple Bayesian Inference Techniques
   - Inference = Integration
   - Integration by Summing
   - Monte Carlo Integration
5. Advanced Bayesian Inference Techniques
   - Markov Chain Monte Carlo Integration
   - Laplace Approximation
   - Variational Approximation
   - Others (Message Passing Algorithms)
6. Survey of Popular Models
   - Latent Dirichlet Allocation
   - Integrating Topics and Syntax
   - Matching Words and Pictures
7. Pointers to Literature on Other Topics
8. Conclusions

2.2 Target Audience

This tutorial should be accessible to anyone with a basic understanding of statistics (familiarity with EM would help, but is not necessary). I use a query-focused summarization task as a motivating running example for the tutorial, which should be of interest to researchers in natural language processing and in information retrieval.

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 301–302,
New York City, June 2006.

Hal's research interests lie at the intersection of machine learning and natural language processing. He works primarily on problems in automatic document summarization and information extraction, using a variety of machine learning techniques. As a Bayesian, he has successfully applied variational inference and expectation propagation techniques to unsupervised learning problems in summarization. He has also successfully applied nonparametric infinite Bayesian models to problems in supervised clustering. In December 2005, he co-organized (with Yee Whye Teh, National University of Singapore) a workshop on "Bayesian Methods for NLP" at the Conference for Neural Information Processing Systems.

3. Graph-Based Algorithms For Natural Language Processing And Information Retrieval

Rada Mihalcea, University of North Texas, and Dragomir Radev, University of Michigan

Graph theory is a well studied discipline, and so are the fields of natural language processing and information retrieval. However, most of the times, they are perceived as different disciplines, with different algorithms, different applications, and different potential end-users.

The goal of this tutorial is to provide an overview of methods and applications in natural language processing and information retrieval that rely on graph-based algorithms. This will include techniques for graph traversal, minimum path length, min-cut algorithms, minimum spanning trees, random walks, etc. and their application to information retrieval and Web search, text understanding (word sense disambiguation and semantic classes), parsing, text summarization, keyword extraction, text clustering, and others.

3.1 Tutorial Outline

1. Graph-based Algorithms Basics
    * Vectors, matrices, graphs
    * Graph representations and notations
    – Traversal, min-cut/max-flow, matching
    * Algorithms for graph traversal
    * Minimum path length
    * Minimum spanning trees
    * Min-cut/max-flow algorithms
    * Graph-matching algorithms
    – Ranking, clustering, learning
    * Eigenvector analysis
    * Node-ranking algorithms
    * Graph-based centrality
    * Graph-based clustering
    * Machine learning on graphs
2. Information Retrieval applications
    * Web-page ranking
    * Text classification and clustering
3. Natural language processing applications
    – Semantics
    * Word sense disambiguation
    * Semantic classes
    * Textual entailment
    * Sentiment classification
    – Syntax, Summarization
    * Dependency parsing
    * Prepositional attachment
    * Keyword extraction
    * Text summarization

3.2 Target Audience

This tutorial is intended for researchers and practitioners who seek a general understanding of the application of graph-theoretical representations and algorithms to natural language processing and information

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 303–304,
New York City, June 2006.

retrieval. It is introductory in nature, no special knowledge or background is required.

Rada Mihalcea is an Assistant Professor of Computer Science at the University of North Texas. Her research interests are in lexical semantics, graph-based algorithms for natural language processing and information retrieval, minimally supervised natural language learning, and multilingual natural language processing. She has published more than 80 articles in books, journals, and proceedings, in these and related areas. 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 serves on the editorial board of the journal of Computational Linguistics, the journal of Language Resources and Evaluations, and the recently established journal of Interesting Negative Results in Natural Language Processing and Machine Learning. Her research is supported by NSF, Google, and the state of Texas.

Dragomir Radev is an Associate Professor of Information, of Computer Science and Engineering, and of Linguistics at the University of Michigan. He has a PhD in Computer Science from Columbia University. He has held numerous posts within NAACL and ACL. He is on the editorial boards of Information Retrieval and the Journal of Artificial Intelligence Research and was recently nominated to the board of the Journal of Natural Language Engineering. He has co-chaired 5 ACL/NAACL workshops and given 6 tutorials at venues like SIGIR, AAAI, and RANLP. Dragomir's current interests are in text summarization, information extraction, information retrieval, graph models, semi-supervised learning, and machine translation. He has more than 50 peer-reviewed papers as well as more than 50 talks at various universities and other venues. Dragomir's work has been funded by NSF, NIH, and ONR.

4. Automatic Spoken Document Processing for Retrieval and Browsing

Ciprian Chelba, Google, and T. J. Hazen, MIT

Ever increasing computing power and connectivity bandwidth together with falling storage costs is resulting in overwhelming amounts of multimedia data being produced, exchanged, and stored. One key application area in this realm is the search and retrieval of spoken audio documents. As storage becomes cheaper, the availability and usefulness of large collections of spoken documents is limited strictly by the lack of adequate technology to exploit them. Manually transcribing speech is expensive and sometimes outright impossible due to privacy concerns. This leads us to exploring an automatic approach to searching and navigating spoken document collections. This tutorial will present an overview of speech transcription, indexing, and search technologies for spoken documents, with an emphasis on a corpus containing recorded academic lectures. The tutorial will point out general problems in this area and suggest possible solutions. Included in the tutorial will be a discussion of scenarios and previous projects in the area of spoken document retrieval, issues of automatic transcription of long audio files, and techniques for the indexing and retrieval of spoken audio files.

4.1 Tutorial Outline

1. Introduction: Scenarios/Previous Work/Corpora
   - Scenarios:
     * Economic considerations for viability of such technology
     * Scenarios where technology is not expected to be useful
     * Scenarios where technology is expected to be useful:
   - Broadcast News
     * Characteristics
     * Meta-data annotation
     * Past work (HP SpeechBot, BBN, TREC, PodZinger, etc.)
   - Academic & Scientific Lectures
     * Examples (OCW, CSJ, MICASE)
     * Characteristics
     * Challenges and opportunities
2. Automatic Speech Transcription
   - Overview of speech recognition models and processing
   - Vocabulary Issues
     * Examination of vocabulary statistics and coverage
     * Vocabulary expansion from supplemental materials
   - Language Modeling Issues
     * Spontaneous conversational speech vs. read speech
     * Appropriateness of written materials
     * Language model adaptation
   - Acoustic Modeling Issues
     * Speaker independent modeling
     * Speaker dependent modeling
     * Supervised and unsupervised adaptation
   - Out-Of-Vocabulary (OOV) modeling
     * Methods for recognizing OOV words
     * Phonetic transcription of OOV words

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 305–306,
New York City, June 2006.

3. Audio Retrieval
   - Overview of text retrieval algorithms:
     * TF-IDF/vector space methods
     * Probabilistic methods
     * Large scale web search (Google)
     * Inverted indexing; query processing/language.
   - Speech recognition lattices:
     * Word/phone/OOV-models for generation
     * Lattice accuracy vs. 1-best accuracy
   - Query processing (OOV problem)/language:
     * ”Soft”-indexing with pruning to control size
     * Combine sub-word and word-level indexing/recognition results
   - Relevance scoring:
     * Proximity
     * Incorporating multiple data streams: speech, text, title, author, abstract, etc.
     * Tuning precision/recall at query run-time
   - Evaluation:
     * Basic Metrics: Precision/Recall
     * Ordered list metrics: Kendall-Tau, Spearman
     * TREC measures and package (Mean Average Precision, R-precision)
     * Issues with evaluating speech data
   - User interface:
     * Issues in consuming speech (as opposed to text, images)
     * Pro’s/con’s for displaying transcription
     * Navigation in long documents with errorful transcriptions
     * Segmentation: topic boundaries, keywords, summaries

4.2 Target Audience

This tutorial is designed for people interested in learning about the technologies used to transcribe, process, search, and retreive spoken audio materials. Detailed prior knowledge of speech recognition and/or search technologies is not required.

Ciprian Chelba is a Research Scientist with Google. Previously he worked as a Researcher in the Speech Technology Group at Microsoft Research. His core research interests are in statistical modeling of natural language and speech. Recent projects include speech content indexing for search in spoken documents, discriminative language modeling for large vocabulary speech recognition, as well as speech and text classification.

Timothy J. Hazen is a Research Scientist at the MIT Computer Science and Artificial Intelligence Laboratory where he works in the areas of automatic speech recognition, automatic person identification, multi-modal speech processing, and conversational speech systems. For the last two years he has been a key contributor to the MIT Spoken Lecture Processing Project.

5. Inductive Semi-supervised Learning Methods for Natural Language Processing

Anoop Sarkar and Gholamreza Haffari, Simon Fraser University

Supervised machine learning methods which learn from labelled (or annotated) data are now widely used in many different areas of Computational Linguistics and Natural Language Processing. There are widespread data annotation endeavours but they face problems: there are a large number of languages and annotation is expensive, while at the same time raw text data is plentiful. Semi-supervised learning methods aim to close this gap.

The last 6-7 years have seen a surge of interest in semi-supervised methods in the machine learning and NLP communities focused on the one hand on analysing the situations in which unlabelled data can be useful, and on the other hand, providing feasible learning algorithms.

This recent research has resulted in a wide variety of interesting methods which are different with respect to the assumptions they make about the learning task. In this tutorial, we survey recent semi-supervised learning methods, discuss assumptions behind various approaches, and show how some of these methods have been applied to NLP tasks.

5.1 Tutorial Outline

1. Introduction
   - Spectrum of fully supervised to unsupervised learning, clustering vs. classifiers or model-based learning
   - Inductive vs. Transductive learning
   - Generative vs. Discriminative learning
2. Mixtures of Generative Models
   - Analysis
   - Stable Mixing of Labelled and Unlabelled data
   - Text Classification by EM
3. Multiple view Learning
   - Co-training algorithm
   - Yarowsky algorithm
   - Co-EM algorithm
   - Co-Boost algorithm
   - Agreement Boost algorithm
   - Multi-task Learning
4. Semi-supervised Learning for Structured Labels (Discriminative models)
   - Simple case: Random Walk
   - Potential extension to Structured SVM
5. NLP tasks and semi-supervised learning
   - Using EM-based methods to combine labelled and unlabelled data
   - When does it work? Some negative examples of semi-supervised learning in NLP
   - Examples of various NLP tasks amenable to semi-supervised learning: chunking, parsing, word-sense disambiguation, etc.
   - Semi-supervised methods proposed within NLP and their relation to machine learning methods covered in this tutorial
   - Semi-supervised learning for structured models relevant for NLP such as sequence learning and parsing
   - Semi-supervised learning for domain adaptation in NLP

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 307–308,
New York City, June 2006.

5.2 Target Audience

The target audience is expected to be researchers in computational linguistics and natural language processing who wish to explore methods that will possibly allow learning from smaller size labelled datasets by exploiting unlabelled data. In particular those who are interested in NLP research into new languages or domains for which resources do not currently exist, or in novel NLP tasks that do not have existing large amounts of annotated data. We assume some familiarity with commonly used supervised learning methods in NLP.

Anoop Sarkar is an Assistant Professor in the School of Computing Science at Simon Fraser University. His research has been focused on machine learning algorithms applied to the study of natural language. He is especially interested in algorithms that combine labeled and unlabeled data and learn new information with weak supervision. Anoop received his PhD from the Department of Computer and Information Science at the University of Pennsylvania, with Prof. Aravind Joshi was his advisor. His PhD dissertation was entitled Combining Labeled and Unlabeled Data in Statistical Natural Language Parsing. A full list of papers is available at http://www.cs.sfu.ca/ anoop. His email address is anoop@cs.sfu.ca

Gholamreza Haffari is a second year PhD student in the School of Computing Science at Simon Fraser University. He is working under the supervision of Prof. Sarkar towards a thesis on semi-supervised learning for structured models in NLP. His home page is http://www.cs.sfu.ca/ ghaffar1, and his email address is ghaffar1@cs.sfu.ca

6. Automatic Semantic Role Labeling

Scott Wen-tau Yih and Kristina Toutanova, Microsoft Research

The goal of semantic role labeling is to map sentences to domain-independent semantic representations, which abstract away from syntactic structure and are important for deep NLP tasks such as question answering, textual entailment, and complex information extraction. Semantic role labeling has recently received significant interest in the natural language processing community. In this tutorial, we will first describe the problem and history of semantic role labeling, and introduce existing corpora and other related tasks. Next, we will provide a detailed survey of state-of-the-art machine learning approaches to building a semantic role labeling system. Finally, we will conclude the tutorial by discussing directions for improving semantic role labeling systems and their application to other natural language problems.

6.1 Tutorial Outline

1. Introduction
   - What is semantic role labeling?
   - Why is SRL important?
   - Existing corpora: FrameNet & PropBank
   - Corpora in development
   - Relation to other tasks
2. Survey of Existing SRL Systems
   - History of the development of automatic SRL systems
   - Pioneering Work
   - Basic architecture of a generic SRL system
   - Major components
   - Machine learning technologies
   - CoNLL-04 and CoNLL-05 shared tasks on SRL
   - Details of several CoNLL-05 systems
   - Overall comparisons of CoNLL-05 systems
3. Analysis of Systems and Future Directions
   - Error Analysis
   - Influence of parser errors
   - Per argument performance
   - Directions for improving SRL
4. Applications
   - Information Extraction
   - Textual Entailment
   - Machine Translation

6.2 Target Audience

The main target audience is NLP students and researchers who are interested in learning about semantic role labeling, but have not followed all developments in the field. Additionally, researchers already working on semantic role labeling should profit from a global view and summary of relevant work. The tutorial will also be valuable for researchers working in the related areas of information extraction and spoken language understanding.

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 309–310,
New York City, June 2006.

Scott Wen-tau Yih received his PhD in Computer Science from the University of Illinois at Urbana-Champaign in 2005 and is currently a Post-Doc Researcher in the Machine Learning and Applied Statistics group at Microsoft Research. His research focuses on different problems in natural language processing and machine learning, such as information extraction and semantic parsing. Scott has published several papers on semantic role labeling in CoNLL-04&05, COLING-04 and IJCAI-05. The SRL system he built at UIUC was the best system in the CoNLL-05 shared task.

Kristina Toutanova obtained her PhD in Computer Science from Stanford University in 2005 and joined Microsoft Research as a Researcher in the Natural Language Processing group. Her areas of expertise include semantic role labeling, syntactic parsing, machine learning, and machine translation. Kristina has published two papers on semantic role labeling in CoNLL-05 and ACL-05. The SRL system she built at Stanford was the runner-up system in the CoNLL-05 shared task.