NAACL HLT 2007

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

Tutorial Abstracts

Marti Hearst, Gina-Anne Levow, and James Allan
Tutorial Chairs

22-27 April 2007
Rochester, NY, USA

Table of Contents

Tutorial title: BioNLP

Speaker: K. Bretonnel Cohen

Description of the tutorial topic and content

This tutorial will provide general natural language processing specialists with an introduction to the field of "BioNLP"—natural language processing in the fields of medicine and biology. This field has long roots in the history of natural language processing, but has been an absolutely burgeoning area of interest in recent years. The past few years have been characterized by an unusual mixing of bioinformatics and NLP specialists at the conferences of both communities: ACL or NAACL has now hosted workshops on BioNLP every year since 2002, with excellent attendance numbers, and bioinformatics and medical informatics meetings have featured NLP papers, sessions, and SIG meetings since the late 1990s. Recent MUC-like and TREC-sponsored shared tasks have had some unusual results, and the implications of these findings should make for an interesting tutorial for the general NLP specialist.

BioNLP presents unique challenges in a number of areas, ranging from low-level processing tasks—tokenization and sentence boundary detection are demonstrably different tasks in biomedical publications than in newswire text—to high-level conceptual issues, such as theoretical issues in predicate-argument structure representation, which have been a topic of much discussion in recent work in the field. Despite the many challenges that are unique to biomedical text, most of the sub-topics of NLP are the subject of current research in the BioNLP community—information retrieval, named entity recognition, information extraction, text classification, semantic role labelling, coreference resolution, question-answering, parsing, morphological analysis, and discourse analysis. Thus, there are interesting challenges in BioNLP for almost anyone working in natural language processing.

One unique advantage to the field of BioNLP is the wide availability of resources, including an enormous body of freely available text. The tutorial will include an overview of a variety of publicly available BioNLP resources, including:

- A variety of domain-specific ontologies, including the popular Gene Ontology
- Corpora, including the popular GENIA corpus and a number of less-well-known but valuable corpora and text collections, some of them featuring full text

One potential stumbling block in the field of BioNLP is the requirement for domain knowledge. The tutorial will include a brief overview of just enough biology to enable the NLP specialist to comprehend the topics under discussion in typical biomedical texts, if not enough to follow the specifics of the discussion.

The core of the tutorial will be an overview of current "hot topics" in BioNLP, including:

- Recent shared tasks and their results
- High-arity relations in information extraction, semantic representation, and semantic role labelling
- Modelling certainty and speculation
- Portability from general-domain to domain-specific tasks, and between sub-domains

Brief outline of the tutorial structure

1. Just enough biology for BioNLP
- genes, proteins, and cells

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL pages 1–2,
Rochester, April 2007.

- genotypes, phenotypes, and high-level structures

2. Why bioscientists fund and publish research in BioNLP
- clinical versus genomic NLP
- the genomic and high-throughput revolutions in biology
- three different biomedical user types and their different text mining needs

3. Some things that make BioNLP different
- issues in tokenization, semantic normalization, word sense disambiguation, and named entity recognition, and how to approach them successfully
- special challenges in biomedical corpus construction

4. Getting up to speed: 10 essential papers and resources that will allow you to read most other papers in BioNLP
- Fukuda, Collier, and Hirschman on named entity recognition
- Blaschke, Craven, and Friedman on information extraction
- The Gene Ontology, Entrez Gene, the GENIA corpus, and the PubMed/MEDLINE database

5. Shared tasks: recent MUC-like and TREC evaluations and their sometimes surprising results
- the KDD Cup in 2002
- JNLPBA in 2004
- BioCreative 2004 & 2006
- the TREC Genomics track, 2003-2006

6. Current hot topics in BioNLP:
- the right model for biomedical semantic representation
- modelling certainty, speculation, and other aspects of discourse structure in scientific texts
- trends in Open Access publishing and their implications for BioNLP: issues in dealing with "full text"
- portability of tools, grammars, and resources
- true integration of NLP into laboratory data interpretation

Intended audience

The intended audience is general NLP specialists. No prior background in the biomedical domain is assumed.

Speaker bio

Kevin Bretonnel Cohen
kevin.cohen@gmail.com
http://compbio.uchsc.edu/Hunter_lab/Cohen
Phone number: 303-916-2417

Kevin leads the Biomedical Text Mining Group at the Center for Computational Pharmacology in the University of Colorado School of Medicine. He has been involved in biomedical NLP in the industrial and academic worlds since 1997. He has worked in both the clinical and the genomic fields, on technologies including information extraction, corpus construction, statistical language modelling for speech recognition, named entity recognition, and computational lexical semantics. He has organized several workshops and conference sessions on BioNLP at ACL, NAACL, and bioinformatics meetings, and has presented tutorials on BioNLP for non-NLP specialists at the Pacific Symposium on Biocomputing, the University of Colorado at Denver Center for Computational Biology, and (this spring) at the Medical Library Association Annual Meeting.

Statistical Language Models for Information Retrieval

ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana_Champaign

Statistical language models have recently been successfully applied to many information retrieval problems. A great deal of recent work has shown that statistical language models not only achieve superior empirical performance, but also facilitate parameter tuning and provide a more principled way, in general, for modeling various kinds of complex and non-traditional retrieval problems.

The purpose of this tutorial is to systematically review the recent progress in applying statistical language models to information retrieval with an emphasis on the underlying principles and framework, empirically effective language models, and language models developed for non-traditional retrieval tasks. Tutorial attendees can expect to learn the major principles and methods of applying statistical language models to information retrieval, the outstanding problems in this area, as well as obtain comprehensive pointers to the research literature. The tutorial should appeal to both people working on information retrieval with an interest in applying more advanced language models and those who have a background on statistical language models and wish to apply them to information retrieval. Attendees will be assumed to know basic probability and statistics.

The outline of the tutorial is as follows:

1. Introduction
   (a) Information Retrieval (IR)
   (b) Statistical Language Models (SLMs)
   (c) Applications of SLMs to IR
2. The Basic Language Modeling Approach
   (a) Query likelihood document ranking
   (b) Smoothing of language models
   (c) Why does it work?
   (d) Variants of the basic LM
3. More Advanced Language Models
   (a) Improving the basic LM approach
   (b) Feedback and alternative ways of using LMs
4. Language Models for Special Retrieval Tasks
   (a) Cross-language IR
   (b) Distributed IR

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL pages 3–4,
Rochester, April 2007.

(c) Structured document retrieval
(d) Personalized/context-sensitive retrieval
(e) Expert retrieval
(f) Modeling redundancy
(g) Predicting query difficulty
(h) Subtopic retrieval

5. A General Framework for Applying SLMs to IR
   (a) Risk minimization framework
   (b) Special cases
   (c) Generative relevance hypothesis

6. Summary
   (a) SLMs vs. traditional methods: Pros & Cons
   (b) What we have achieved so far
   (c) Challenges and future directions

ChengXiang Zhai is an Assistant Professor of Computer Science at the University of Illinois at Urbana-Champaign, where he also holds a joint appointment at the Institute for Genomic Biology and the Graduate School of Library and Information Science. He received a Ph.D. in Computer Science from Nanjing University in 1990, and a Ph.D. in Language and Information Technologies from Carnegie Mellon University in 2002. He worked at Clairvoyance Corp. as a Research Scientist and, later, a Senior Research Scientist from 1997 to 2000. His research interests include information retrieval, text mining, natural language processing, machine learning, and bioinformatics. He serves on the editorial board of ACM Transactions on Information Systems, and is the program co-chair of ACM CIKM 2004 and NAACL HLT 2007. He is an invited participant of the National Academy of Engineering's 2006 US Frontiers of Engineering Symposium. He received an NSF CAREER Award in 2004, the ACM SIGIR 2004 Best Paper Award, and the 2004 Presidential Early Career Award for Scientists and Engineers (PECASE).

Arabic Dialect Processing Tutorial

Mona Diab and **Nizar Habash**
Center for Computational Learning Systems
Columbia University
{mdiab,habash}@cs.columbia.edu

Language exists in a natural continuum, both historically and geographically. The term *language* as opposed to *dialect* is only an expression of power and dominance of one group/ideology over another. In the Arab world, politics (Arab nationalism) and religion (Islam) are what shape the perception of the distinction between *the* Arabic language and an Arabic dialect. This power relationship is similar to others that exist between languages and their dialects. However, the high degree of difference between standard Arabic and its dialects and the fact that standard Arabic is not any Arab's native language sets the Arabic linguistic situation apart.

As such, the Arabic language can be conceived of as a collection of multiple variants among which Modern Standard Arabic (MSA) has a special status as the formal written standard language of the media, culture and education across the Arab world. The other variants are informal spoken dialects that are the mediums of communication for daily life. Arabic dialects substantially differ from MSA and each other in terms of phonology, morphology, lexical choice and syntax.

MSA is the official language of the Arab world. It is the primary language of the media and culture. MSA is syntactically, morphologically and phonologically based on Classical Arabic, the language of the Qur'an (Islam's Holy Book). Lexically, however, it is much more modern. It is not a native language of any Arabs but is the language of education across the Arab world. MSA is primarily written not spoken.

The Arabic dialects, in contrast, are the true native language forms. They are generally restricted in use for informal daily communication. They are not taught in schools or even standardized although there is a rich popular dialect culture of folktales, songs, movies, and TV shows. Dialects are primarily spoken not written. However this is quite changing since more Arabs are gaining access to electronic media of communication such as emails and newsgroups. Arabic dialects are loosely related to Classical Arabic. They are the result of the interaction between different ancient dialects of Classical Arabic and the indigenous languages that existed in today's Arab world together with influences from colonization and interaction with neighboring countries. For example, Algerian Arabic has a lot of influences from it's ancient indigenous language Berber as well as French due to the French occupation.

Arabic dialects vary on many dimensions – primarily, geography and social class. Geolinguistically, the Arab world can be divided in many different ways. The following is only one of many: **Levantine Arabic** includes the dialects of Lebanon, Syria, Jordan, Palestine and Israel. **Gulf Arabic** includes the dialects of Kuwait, Saudi Arabia, United Arab Emirates, Bahrain, and Qatar. Iraqi and Omani Arabic are included some times. **Egyptian Arabic** covers the dialects of the Nile valley: Egypt and Sudan. **North African Arabic** covers the dialects of Morocco, Algeria, Tunisia and Mauritania. Libya is sometimes included. **Yemenite Arabic** is often considered its own class. **Maltese Arabic** is not always considered an Arabic dialect. It is the only Arabic variant that is considered a separate language and is written with Latin script.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL pages 5–6,
Rochester, April 2007.

Socially, it is common to distinguish three sub-dialects within each dialect region: city dwellers, peasants/farmers and Bedouins. The three degrees are often associated with a class hierarchy in which rich settled city dwellers are on top and Bedouins are on bottom. Different social associations exist as common in many other languages around the world. For example, the city dialect is considered less marked, better and smarter; whereas the Bedouin dialect is considered lower class, rough, yet pure to the origin of the language.

The relationship between MSA and the dialect in a *specific region* is rather complex. Arabs do not think of these two as separate languages. This particular perception leads to a special kind of coexistence between two forms of language that serve different purposes. This kind of situation is what linguists term *diglossia*. Although the two variants have clear domains of prevalence: formal written (MSA) versus informal spoken (dialect), there is a large gray area in between and it is often filled with mixing of the two forms.

For Natural Language Processing (NLP), the existence of dialects for any language constitutes a challenge in general since it adds another set of variation dimensions from a known standard. The problem is particularly interesting and challenging in Arabic and its different dialects, where the diversion from the standard could, in some linguistic theories, warrant a classification as a different language. This problem would not be as pronounced if standard Arabic were to be a living language, however it is not. Any realistic and practical approach to processing Arabic will have to account for dialectal usage since it is so pervasive.

In this tutorial, we highlight different dialectal phenomena. We discuss how dialects migrate from the standard and why they pose challenges to NLP. Our tutorial will have four different parts: First, we describe a background layout of issues for standard Arabic NLP. Then we discuss a high level generic view of dialects and their different aspects that are of interest for the NLP community. We address both text and speech issues in addition to standardization issues. We focus in depth on two aspects of dialect processing in the third and fourth parts of the tutorial, namely, dialectal morphology and dialectal syntactic parsing. Throughout the presentation we will make references to the different resources available and draw contrastive links with standard Arabic and English. Moreover, we will discuss annotation standards as exemplified in the Linguistic Data Consortium Arabic Treebank. We will provide links to recent publications and available toolkits/resources for all four sections.

This tutorial is designed for computer scientists and linguists alike. No knowledge of Arabic is required. However, we recommend taking a look at Nizar Habash's Arabic NLP tutorial[1] which will be reviewed in the first quarter of the tutorial.

[1] http://www.ccls.columbia.edu/cadim/presentations.html

Introduction to Classification: Likelihoods, Margins, Features, and Kernels

Tutorial for NAACL-HLT 2007

Dan Klein
Computer Science Division
University of California, Berkeley
klein@cs.berkeley.edu

Overview

Statistical methods in NLP have exploited a variety of classification techniques as core building blocks for complex models and pipelines. In this tutorial, we will survey the basic techniques behind classification. We first consider the basic principles, including the principles of maximum likelihood and maximum margin. We then discuss several core classification technologies: naive Bayes, perceptrons, logistic regression, and support vector machines. The discussion will include the key optimization ideas behind their training and the empirical trade-offs between the various classifiers. Finally, we consider the extension to kernels and kernelized classification: what can kernels offer and what is their cost? The presentation is targeted to NLP researchers new to these methods or those wanting to understand more about how these techniques are interconnected.

Topics

1. Basics of classification
   (a) Feature-based representations
   (b) Linear classifiers
   (c) Principles of classification: likelihood and margin
   (d) Smoothing and regularization
   (e) Structured classification
2. Specific techniques
   (a) Perceptrons
   (b) Naive Bayes
   (c) Logistic regression / maximum entropy
   (d) Support vector machines
   (e) Comparison and trade-offs
3. Kernel methods
   (a) Why kernels (and why not)?
   (b) Kernelized linear classifiers
   (c) Kernelized perceptrons
   (d) Kernelizing SVMs and logistic regression
   (e) Advanced kernels and structure

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL page 7,
Rochester, April 2007.

Author Index