ACL-08: HLT

46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies

Tutorial Abstracts

June 15–20, 2008
The Ohio State University
Columbus, Ohio, USA

Production and Manufacturing by
Omnipress Inc.
2600 Anderson Street
Madison, WI 53707
USA

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

Tutorials Chairs:

Ani Nenkova, University of Pennsylvania, USA
Marilyn Walker, University of Sheffield, UK
Eugene Agichtein, Emory University, USA

Table of Contents

Introduction to Computational Advertising
Evgeniy Gabrilovich, Vanja Josifovski and Bo Pang . . . 1

Building Practical Spoken Dialog Systems
Antoine Raux, Brian Langner, Alan W. Black and Maxine Eskenazi . . . 2

Semi-Supervised Learning for Natural Language Processing
John Blitzer and Xiaojin Jerry Zhu . . . 3

Advanced Online Learning for Natural Language Processing
Koby Crammer . . . 4

Speech Technology: From Research to the Industry of Human-Machine Communication
Roberto Pieraccini . . . 5

Interactive Visualization for Computational Linguistics
Christopher Collins, Gerald Penn and Sheelagh Carpendale . . . 6

Tutorial Program

Sunday, June 15, 2008

Morning Tutorials

9:00–12:30 *Introduction to Computational Advertising*
Evgeniy Gabrilovich, Vanja Josifovski and Bo Pang

9:00–12:30 *Building Practical Spoken Dialog Systems*
Antoine Raux, Brian Langner, Alan W. Black and Maxine Eskenazi

9:00–12:30 *Semi-Supervised Learning for Natural Language Processing*
John Blitzer and Xiaojin Jerry Zhu

Afternoon Tutorials

2:00–5:30 *Advanced Online Learning for Natural Language Processing*
Koby Crammer

2:00–5:30 *Speech Technology: From Research to the Industry of Human-Machine Communication*
Roberto Pieraccini

2:00–5:30 *Interactive Visualization for Computational Linguistics*
Christopher Collins, Gerald Penn and Sheelagh Carpendale

Introduction to Computational Advertising

Evgeniy Gabrilovich **Vanja Josifovski** **Bo Pang**
Yahoo! Research
701 First Avenue
Sunnyvale, CA 94085, USA
{gabr,vanjaj,bopang}@yahoo-inc.com

1 Introduction

Web advertising is the primary driving force behind many Web activities, including Internet search as well as publishing of online content by third-party providers. Even though the notion of online advertising barely existed a decade ago, the topic is so complex that it attracts attention of a variety of established scientific disciplines, including computational linguistics, computer science, economics, psychology, and sociology, to name but a few. Consequently, a new discipline — Computational Advertising — has emerged, which studies the process of advertising on the Internet from a variety of angles. A successful advertising campaign should be relevant to the immediate user's information need as well as more generally to user's background and personalized interest profile, be economically worthwhile to the advertiser and the intermediaries (e.g., the search engine), as well as be aesthetically pleasant and not detrimental to user experience.

2 Content Overview

In this tutorial, we focus on one important aspect of online advertising that is relevant to the ACL and HLT communities, namely, contextual relevance. There are two main scenarios for online advertising, as advertisers might request to display their ads for a query submitted to a Web search engine, or for a Web page that the user reads online.The former scenario is called sponsored search, since ads are matched to the Web search results, and the latter — content match, as ads are matched to a larger amount of content. It is essential to emphasize that in both cases the context of user actions is defined by a body of text, which could be quite short in the case of sponsored search or fairly long in the case of content match. Consequently, the ad matching problem lends itself to many NLP methods, but also poses numerous challenges and open research problems in text summarization, natural language generation, named entity extraction, computer-human interaction, and others.

At first approximation, the process of obtaining relevant ads can be reduced to conventional information retrieval, where we construct a query that describes the user's context, and then execute this query against a large inverted index of ads. We show how to augment the standard information retrieval approach using query expansion and text classification techniques. First, we demonstrate how to employ a relevance feedback assumption and use Web search results produced by the query. We also go beyond the conventional bag of words indexing, and construct additional features by classifying both the input context and the ad descriptions with respect to a large external taxonomy. A third type of features is constructed from a lexicon of named entities obtained by analyzing the entire Web as a corpus.

We present a unified approach to Web advertising, which uses the same underlying infrastructure to handle both sponsored search and content match scenarios. The last part of the tutorial will be devoted to recent research results as well as open problems, such as automatically classifying cases when no ads should be shown, handling geographic names (and more generally, location awareness), and context modeling for vertical portals.

Tutorial Abstracts of ACL-08: HLT, page 1,
Columbus, Ohio, USA, June 2008.

Building Practical Spoken Dialog Systems

Antoine Raux[1], Brian Langner[2], Alan W Black[3], Maxine Eskenazi[4]
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
{antoine,blangner,awb,max}@cs.cmu.edu

1 Abstract

This tutorial will give a practical description of the free software Carnegie Mellon Olympus 2 Spoken Dialog Architecture. Building real working dialog systems that are robust enough for the general public to use is difficult. Most frequently, the functionality of the conversations is severely limited - down to simple question-answer pairs. While off-the-shelf toolkits help the development of such simple systems, they do not support more advanced, natural dialogs nor do they offer the transparency and flexibility required by computational linguistic researchers. However, Olympus 2 offers a complete dialog system with automatic speech recognition (Sphinx) and synthesis (SAPI, Festival) and has been used, along with previous versions of Olympus, for teaching and research at Carnegie Mellon and elsewhere for some 5 years. Overall, a dozen dialog systems have been built using various versions of Olympus, handling tasks ranging from providing bus schedule information to guidance through maintenance procedures for complex machinery, to personal calendar management. In addition to simplifying the development of dialog systems, Olympus provides a transparent platform for teaching and conducting research on all aspects of dialog systems, including speech recognition and synthesis, natural language understanding and generation, and dialog and interaction management.

The tutorial will give a brief introduction to spoken dialog systems before going into detail about how to create your own dialog system within Olympus 2, using the Let's Go bus information system as an example. Further, we will provide guidelines on how to use an actual deployed spoken dialog system such as Let's Go to validate research results in the real world. As a possible testbed for such research, we will describe Let's Go Lab, which provides access to both the Let's Go system and its genuine user population for research experiments.

2 Outline

Part 1

- 1.1 Introduction
- 1.2 Overview of current spoken dialog system architectures
- 1.3 Description of the Olympus2 dialog architecture
- 1.4 How to build an Olympus2 spoken dialog system

Part 2

- 2.1 Advanced Topics
  - a. Improving ASR
  - b. Improving TTS
  - c. Dealing with ASR Errors
  - d. Logs and Tools
- 2.2 Using Olympus2 for research and applications
- 2.3 Final summary

[1] http://www.cs.cmu.edu/~antoine
[2] http://www.cs.cmu.edu/~blangner
[3] http://www.cs.cmu.edu/~awb
[4] http://www.cs.cmu.edu/~max

Tutorial Abstracts of ACL-08: HLT, page 2,
Columbus, Ohio, USA, June 2008.

Semi-supervised Learning for Natural Language Processing

John Blitzer
Natural Language Computing Group
Microsoft Research Asia
Beijing, China
blitzer@cis.upenn.edu

Xiaojin Jerry Zhu
Department of Computer Science
University of Wisconsin, Madison
Madison, WI, USA
jerryzhu@cs.wisc.edu

1 Introduction

The amount of unlabeled linguistic data available to us is much larger and growing much faster than the amount of labeled data. Semi-supervised learning algorithms combine unlabeled data with a small labeled training set to train better models. This tutorial emphasizes practical applications of semi-supervised learning; we treat semi-supervised learning methods as tools for building effective models from limited training data. An attendee will leave our tutorial with

1. A basic knowledge of the most common classes of semi-supervised learning algorithms and where they have been used in NLP before.

2. The ability to decide which class will be useful in her research.

3. Suggestions against potential pitfalls in semi-supervised learning.

2 Content Overview

Self-training methods Self-training methods use the labeled data to train an initial model and then use that model to label the unlabeled data and retrain a new model. We will examine in detail the co-training method of Blum and Mitchell [2], including the assumptions it makes, and two applications of co-training to NLP data. Another popular self-training method treats the labels of the unlabeled data as hidden and estimates a single model from labeled and unlabeled data. We explore new methods in this framework that make use of declarative linguistic side information to constrain the solutions found using unlabeled data [3].

Graph regularization methods Graph regularization methods build models based on a graph on instances, where edges in the graph indicate similarity. The regularization constraint is one of smoothness along this graph. We wish to find models that perform well on the training data, but we also regularize so that unlabeled nodes which are similar according to the graph have similar labels. For this section, we focus in detail on the Gaussian fields method of Zhu et al. [4].

Structural learning Structural learning [1] uses unlabeled data to find a new, reduced-complexity hypothesis space by exploiting regularities in feature space via unlabeled data. If this new hypothesis space still contains good hypotheses for our supervised learning problem, we may achieve high accuracy with much less training data. The regularities we use come in the form of lexical features that function similarly for prediction. This section will focus on the assumptions behind structural learning, as well as applications to tagging and sentiment analysis.

References

[1] Rie Ando and Tong Zhang. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. JMLR 2005.

[2] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-training. COLT 1998.

[3] Aria Haghighi and Dan Klein. Prototype-driven Learning for Sequence Models. HLT/NAACL 2006.

[4] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised Learning using Gaussian Fields and Harmonic Functions. ICML 2003.

Tutorial Abstracts of ACL-08: HLT, page 3,
Columbus, Ohio, USA, June 2008.

Advanced Online Learning for Natural Language Processing

Koby Crammer
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
crammer@cis.upenn.edu

Introduction: Most research in machine learning has been focused on binary classification, in which the learned classifier outputs one of two possible answers. Important fundamental questions can be analyzed in terms of binary classification, but real-world natural language processing problems often involve richer output spaces. In this tutorial, we will focus on classifiers with a large number of possible outputs with interesting structure. Notable examples include information retrieval, part-of-speech tagging, NP chucking, parsing, entity extraction, and phoneme recognition.

Our algorithmic framework will be that of online learning, for several reasons. First, online algorithms are in general conceptually simple and easy to implement. In particular, online algorithms process one example at a time and thus require little working memory. Second, our example applications have all been treated successfully using online algorithms. Third, the analysis of online algorithms uses simpler mathematical tools than other types of algorithms. Fourth, the online learning framework provides a very general setting which can be applied to a broad setting of problems, where the only machinery assumed is the ability to perform exact inference, which computes a maxima over some score function.

Goals: (1) To provide the audience systematic methods to design, analyze and implement efficiently learning algorithms for their specific complex-output problems: from simple binary classification through multi-class categorization to information extraction, parsing and speech recognition. (2) To introduce new online algorithms which provide state-of-the-art performance in practice backed by interesting theoretical guarantees.

Content: The tutorial is divided into two parts. In the first half we introduce online learning and describe the Perceptron algorithm (Rosenblatt, 1958) and the passive-aggressive framework (Crammer et al., 2006). We then discuss in detail an approach for deriving algorithms for complex natural language processing (Crammer, 2004). In the second half we discuss is detail relevant applications including text classification (Crammer and Singer, 2003), named entity recognition (McDonald et al., 2005), parsing (McDonald, 2006), and other tasks. We also relate the online algorithms to their batch counterparts.

References

K. Crammer and Y. Singer. 2003. A new family of online algorithms for category ranking. *Jornal of Machine Learning Research*, 3:1025–1058.

K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. 2006. Online passive-aggressive algorithms. *JMLR*, 7:551–585.

K. Crammer. 2004. *Online Learning of Complex Categorial Problems*. Ph.D. thesis, Hebrew Universtiy.

R. McDonald, K. Crammer, and F. Pereira. 2005. Flexible text segmentation with structured multilabel classification. In *HLT/EMNLP*.

R. McDonald. 2006. *Discriminative Training and Spanning Tree Algorithms for Dependency Parsing*. Ph.D. thesis, University of Pennsylvania.

F. Rosenblatt. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407.

Tutorial Abstracts of ACL-08: HLT, page 4,
Columbus, Ohio, USA, June 2008.

Speech Technology:
from Research to the Industry of Human-Machine Communication

Roberto Pieraccini
SpeechCycle
26 Broadway, 11th Floor
New York, NY 10004
roberto@speechcycle.com

1 Introduction

This tutorial is about the evolution of speech technology from research to a mature industry. Today, spoken language communication with computers is becoming part of everyday life. Thousands of interactive applications using spoken language technology— known also as "conversational machines"—are only phone calls away, allowing millions of users each day to access information, perform transactions, and get help. Speech recognition, language understanding, text-to-speech synthesis, machine learning, and dialog management enabled this revolution after more than 50 years of research. The industry of speech continues to mature with its evolving standards, platforms, architectures, and business models within different sectors of the market.

2 Content Overview

In this tutorial I will briefly trace the history of speech technology, with a special focus on speech recognition and spoken language understanding, from the early attempts to today's commercial deployments. I will summarily describe the most successful ideas and algorithms that brought to today's technology. I will discuss the struggle for ever increasing performance, the importance of data for training and evaluation, and the role played by government funded projects in creating effective evaluation benchmarks. I will then describe the birth of the speech industry in the mid 1990s, with the role played by the Voice User Interface and dialog engineering disciplines in bringing speech recognition from a laboratory "accuracy challenge" to an enabler of usable interfaces. I will describe the rising of standards (such as VoiceXML, SRGS, SSML, etc.) and their importance in the growth of the market. I will proceed with an overview of the current architectures and processes utilized for creating commercial spoken dialog systems, and will provide several case studies of the use of speech technology. I will conclude with a discussion on the current open problems and challenges.

The tutorial duration will be of about 3 hours with a short break. Several audio and video samples will be shown during the tutorial. The tutorial is directed to a general HLT audience with no prior knowledge of speech technology.

3 Tutorial Outline

- What is speech and why it is difficult to recognize it.
- The history of speech recognition from the early attempts to Hidden Markov Models
- The struggle for performance and the importance of data
- Spoken language understanding and dialog
- The birth of the "spoken dialog" industry
- Industrial standards and architectures
- Case studies
- Open issues and future research

References

Pieraccini, R. Huerta, J., *Where do we go from here? Research and Commercial Spoken Dialog Systems*, Proc. of 6th SIGdial Workshop on Discourse and Dialog, Lisbon, Portugal, 2-3 September, 2005. pp. 1-10

Pieraccini R., Lubensky, D., *Spoken Language Communication with Machines: the Long and Winding Road from research to Business*, in M. Ali and F. Esposito (Eds) : IEA/AIE 2005, LNAI 3533, pp 6-15, 2005, Springer-Verlag..

Tutorial Abstracts of ACL-08: HLT, page 5,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

Interactive Visualization for Computational Linguistics

Christopher Collins and Gerald Penn
Department of Computer Science
University of Toronto
10 King's College Road
Toronto, Ontario, Canada
{ccollins,gpenn}@cs.utoronto.ca

Sheelagh Carpendale
Department of Computer Science
University of Calgary
2500 University Dr. NW
Calgary, Canada
sheelagh@ucalgary.ca

Interactive information visualization is an emerging and powerful research technique that can be used to understand models of language and their abstract representations. Much of what computational linguists fall back upon to improve NLP applications and to model language "understanding" is structure that has, at best, only an indirect attestation in observable data. An important part of our research progress thus depends on our ability to fully investigate, explain, and explore these structures, both empirically and relative to accepted linguistic theory. The sheer complexity of these abstract structures, and the observable patterns on which they are based, usually limits their accessibility — often even to the researchers creating or attempting to learn them.

To aid in this understanding, visual 'externalizations' are used for presentation and explanation — traditional statistical graphs and custom-designed illustrations fill the pages of ACL papers. These visualizations provide *post hoc* insight into the representations and algorithms designed by researchers, but visualization can also assist in the process of research itself. There are special statistical methods, falling under the rubric of "exploratory data analysis," and visualization techniques just for this purpose, in fact, but these are not widely used or even known in CL. These techniques offer the potential for revealing structure and detail in data, before anyone else has noticed them.

When observing natural language engineers at work, we also notice that, even without a formal visualization background, they often create sketches to aid in their understanding and communication of complex structures. These are *ad hoc* visualizations, but they, too, can be extended by taking advantage of current information visualization research.

This tutorial will enable members of the ACL community to leverage information visualization theory into exploratory data analysis, algorithm design, and data presentation techniques for their own research. We draw on fundamental studies in cognitive psychology to introduce 'visual variables' — visual dimensions on which data can be encoded. We also discuss the use of interaction and animation to enhance the usability and usefulness of visualizations.

Topics covered in this tutorial include a review of information visualization techniques that are applicable to CL, pointers to existing visualization tools and programming toolkits, and new directions in visualizing CL data and results. We also discuss the challenges of evaluating visualizations, noting differences from the evaluation methods traditionally used in CL, and discuss some heuristic approaches and techniques used for measuring insight. Information visualizations in CL research can also be measured by the impact they have on algorithm and data structure design.

Information visualization is also filled with opportunities to make more creative visualizations that benefit from the CL community's deeper collective understanding of natural language. Given that most visualizations of language are created by researchers with little or no linguistic expertise, we'll cover some open and very ripe possibilities for improving the state of the art in text-based visualizations.

Tutorial Abstracts of ACL-08: HLT, page 6,
Columbus, Ohio, USA, June 2008.

Author Index

Black, Alan W., 2
Blitzer, John, 3

Carpendale, Sheelagh, 6
Collins, Christopher, 6
Crammer, Koby, 4

Eskenazi, Maxine, 2

Gabrilovich, Evgeniy, 1

Josifovski, Vanja, 1

Langner, Brian, 2

Pang, Bo, 1
Penn, Gerald, 6
Pieraccini, Roberto, 5

Raux, Antoine, 2

Zhu, Xiaojin Jerry, 3