ACL 2007

Tutorial

Abstracts

June 24, 2007

Prague, Czech Republic

Production and Manufacturing by
Omnipress
2600 Anderson Street
Madison, WI 53704
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

Tutorial Chair

Joakim Nivre, Växjö University, Sweden

Tutorials and Presenters

Morning

T2: Usability and Performance Evaluation for Advanced Spoken Dialogue Systems
Kristiina Jokinen and Michael McTear

T4: From Web Content Mining to Natural Language Processing
Bing Liu

T5: Quality Control of Corpus Annotation Through Reliability Measures
Ron Artstein

Afternoon

T1: Bayesian Nonparametric Structured Models
Percy Liang and Dan Klein

T3: Textual Entailment
Ido Dagan, Dan Roth and Fabio Massimo Zanzotto

Usability and Performance Evaluation for Advanced Spoken Dialogue Systems

Kristiina Jokinen
Department of Computer Sciences
University of Tampere / University of Tartu
kristiina.jokinen@helsinki.fi

Michael McTear
School of Computing and Mathematics
University of Ulster at Jordanstown
MF.McTear@ulster.ac.uk

Abstract

The past decade has seen a rapid emergence of dialogue systems that support robust and efficient interaction in spoken natural language. The technology has become mature enough for speech-based interactive applications to be built for practical purposes, and the results are also being applied to novel fields such as situated cognition and robots, embodied conversational agents, meeting assistants, etc. There has also been a growth of research focusing on advanced spoken dialogue systems that aim to increase the system's communicative competence by including aspects of interaction that go beyond the basic techniques of interaction management. Such advanced aspects include e.g. disfluencies, turn-taking, speaker intentions, emotions, multimodality, and adaptation in context.

One of the motivations for furthering the system's interaction capabilities is to improve the system's naturalness and usability in practical applications. However, there are several issues that need to be considered, and requirements and evaluation metrics seem to differ for academic and industrial perspectives. For instance, over the past decade, the research community has focussed on various data-driven methods in contrast to the hand-crafted rules that are used predominantly in commercial systems. The main arguments in support of the data-driven approach concern robust understanding and the assumption that it is ultimately more portable and less labour intensive than hand-crafting, whereas approaches for practical systems usually emphasise rather simple albeit robust solutions, and the importance of adhering to the requirements, needs and preferences of real users.

Research and development presuppose well-defined criteria according to which interactive systems can be evaluated in terms of usability and naturalness. Objective and subjective criteria have been identified and enumerated, and although no consensus has been reached on the general practices, criteria, or metrics of evaluation, it is widely agreed that rigorous evaluation methodologies should be consolidated as part of the research activities, especially when dealing with complex issues as such adaptation, user requirements, and best-practice applications.

This tutorial will focus on methods, problems and challenges in the evaluation of advanced spoken dialogue systems. It is grounded in research that combines various speech and language technology components into an integrated system, and surveys the issues related to the design, evaluation and comparison of such systems. A number of different approaches to robust and efficient interaction management will be reviewed, together with various performance and user evaluation methods in academic and industrial environments. A closer look will be taken at different metrics and usability criteria, as well as automatic design and evaluation methods. Practical requirements for dialogue systems, such as robustness, scalability and portability will also be discussed and exemplified from the point of view of performance evaluation and usability. Special attention will be paid to user evaluation, and to the user's expectations and experience of the system.

References

Jokinen, Kristiina. forthcoming. *Constructive Dialogue Management – Speech Interaction and Rational Agents*. John Wiley & Sons.

Michael McTear. 2004. *Spoken dialogue technology: toward the conversational user interface* Springer Verlag.

From Web Content Mining to Natural Language Progressing

Bing Liu
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan Street, Chicago, IL 60607-7053
liub@cs.uic.edu

Abstract

This tutorial introduces some important tasks of Web content mining and their connections with natural language processing (NLP), and encourages NLP researchers to join the Web content mining research.

1 Introduction

Web mining consists of usage mining, structure mining, and content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web structure mining tries to discover useful knowledge from hyperlinks. Web content mining aims to extract/mine useful knowledge from Web page contents. This tutorial focuses on Web content mining and its extensive connections with natural language processing (NLP).

In the past few years, there was a rapid expansion of activities in Web content mining because of the huge amount of valuable information of almost any imaginable type on the Web and significant economic benefits of such mining. However, due to the heterogeneity and the lack of structure of the Web data, automated discovery of useful knowledge still presents challenging problems. This tutorial introduces several such problems. These problems all have strong connections with NLP. In the tutorial, special attentions will be paid to such connections. Many real-life examples are also given to help participants understand research concepts and see how the technologies may be deployed to real-life applications. The tutorial thus has a mix of research and industry flavor, addressing seminal research ideas and looking at the technology from an industry angle.

2 Tutorial Outline

1. **Structured data extraction**
   The problem
   Wrapper induction
   Automated extraction
   NLP connection
2. **Information integration**
   The problem
   Some integration techniques
   Web query interface integration
   NLP connection
3. **Information synthesis**
   The problem
   Exploiting information redundancy
   Using syntactic patterns
   NLP connection
4. **Opinion mining**
   The abstraction
   Document level sentiment classification
   Sentence level sentiment analysis
   Feature-based opinion mining & summarization
   Comparative sentence and relation mining

3 Presenter Biography

Bing Liu is an associate professor of Computer Science at the University of Illinois at Chicago. His research interests include data, Web and text mining. He has published extensively in these areas in leading conferences, e.g., KDD, WWW, AAAI, IJCAI, ICML and SIGIR. He has served or serves as program chairs, vice chairs or senior PC members of many data mining related conferences, including KDD, WWW, ICDM, SDM, CIKM and PAKDD, and also as an associate editor of IEEE Transactions on Knowledge and Data Engineering, and an associate editor of SIGKDD Explorations.

Quality Control of Corpus Annotation through Reliability Measures

Ron Artstein
Department of Computer Science
University of Essex, Wivenhoe Park
Colchester CO4 3SQ
United Kingdom
artstein@essex.ac.uk

Abstract

The need for quality control of corpus annotation should be obvious: research based on annotated corpora can only be as good as the annotations themselves. In recent years, corpus annotation has expanded from marking basic morphological and syntactic structure to many new kinds of linguistic phenomena. Each new annotation scheme and every individual set of annotation guidelines need to be checked for quality, because quality inferences do not carry over from one scheme to another. A standard way of assessing the quality of an annotation scheme and guidelines is to compare annotations of the same text by two or more independent annotators. While researchers are generally aware of this technique, it seems that the inner workings of the statistics involved and how to interpret them are understood by few. Mechanical application of agreement coefficients found in software packages can lead to serious errors, and it is therefore crucial for people who do research on annotation to become intimately familiar with these statistics.

This tutorial is a thorough introduction to the statistics used for measuring agreement between corpus annotators, and hence for inferring the reliability of the annotation. The tutorial will focus on the mathematics of the various agreement measures, and consequently on the implicit assumptions they make about annotators and annotation errors. A major part of the tutorial will be devoted to agreement coefficients of the kappa family, which are the most commonly used reliability measures in computational linguistics, but the tutorial will also discuss alternative measures such as latent class analysis. The tutorial will not assume advanced mathematical knowledge beyond basic probability theory, and will thus be accessible to most researchers in computational linguistics.

Tutorial outline

1. Motivation
   - Reliability as an indicator of annotation quality in the absence of a test for correctness
   - Agreement between coders as a measure of reliability
2. How to measure agreement
   - Correction for chance agreement (Scott's Pi)
   - Individual coder bias (Cohen's Kappa)
   - Multiple coders (Fleiss)
   - Weighted coefficients (Cohen's weighted Kappa, Krippendorff's Alpha)
3. How to interpret agreement coefficients
   - Using agreement coefficients to infer the proportion of difficult items (Aickin)
   - Inferring error rates on different classes of items (latent class models)
4. Using agreement measures
   - Identifying strong and weak parts of the annotation
   - Adapting the coefficients for specific uses

Bayesian Nonparametric Structured Models

Percy Liang
Computer Science Division
University of California at Berkeley
Berkeley, CA 94720
pliang@cs.berkeley.edu

Dan Klein
Computer Science Division
University of California at Berkeley
Berkeley, CA 94720
klein@cs.berkeley.edu

Abstract

Probabilistic modeling is a dominant approach for both supervised and unsupervised learning in NLP. One constant challenge for models with latent variables is determining the appropriate model complexity, i.e. the question of "how many clusters." While cross-validation can be used to select between a limited number of options, it cannot be feasibly applied in the context of larger hierarchical models where we must balance complexity in many parts of the model at the same time. Nonparametric "infinite" priors such as the *Dirichlet process* are tools from the Bayesian statistics literature which present an elegant solution to this problem and have seen increasing use in recent NLP work. Models based on the Dirichlet process have an infinite number of clusters and rely on the prior to penalize their use, thus allowing the complexity of the model to adapt to the data.

In explaining how to do inference in these new models, we try to dispel two myths: first, that Bayesian methods are too slow and cumbersome, and, second, that Bayesian techniques require a whole new set of algorithmic ideas. We depart from the traditional sampling methodology which has dominated past expositions and focus on *variational inference*, an efficient technique which is a natural extension of EM. This approach allows us to tackle structured models such as HMMs and PCFGs with the benefits of Bayesian nonparametrics while maintaining much of the existing EM machinery so familiar to this community. We provide intuition about the Dirichlet process, variational inference, and their interaction at several different levels. In addition to our foundational presentation, we discuss concrete implementation issues, and demonstrate the empirical properties of these methods on several different models.

Presenter Biography

Percy Liang is a Ph.D. student in computer science at UC Berkeley. He has a BS in math and a BS/MS in computer science from MIT. His research interests include probabilistic modeling for semi-supervised learning in NLP, especially using Bayesian nonparametrics, and approximate inference algorithms for such models. He holds an NSF Graduate Fellowship and a National Defense Science and Engineering Graduate Fellowship.

Dan Klein is an assistant professor of computer science at the University of California, Berkeley (PhD Stanford, MS Oxford, BA Cornell). Professor Klein's research focuses on statistical natural language processing, including unsupervised methods, syntactic parsing, and machine translation. His academic honors include a British Marshall Fellowship, an inaugural Microsoft New Faculty Fellowship, and best paper awards at the ACL, NAACL, and EMNLP conferences.

Textual Entailment

Ido Dagan
Computer Science
Bar-Ilan University
Ramat Gan 52900, Israel
dagan@mac.biu.ac.il

Dan Roth
Computer Science
University of Illinois
Urbana, IL 61801 USA
danr@uiuc.edu

Fabio Massimo Zanzotto
DISP
University of Rome "Tor Vergat"
00133 Rome, Italy
zanzotto@info.uniroma2.it

Abstract

Recognizing Textual Entailment is the task of determining, for example, that the sentence: "Google files for its long awaited IPO" entails that "Google goes public". Determining whether the meaning of a given text passage entails that of another or whether they have the same meaning is a fundamental problem in natural language understanding that requires the ability to abstract over the inherent syntactic and semantic variability in natural language. This challenge is at the heart of many natural language understanding tasks including Question Answering, Information Retrieval and Extraction, Machine Translation, and others that attempt to reason about and capture the meaning of linguistic expressions. The task has attracted significant interest over the last couple of years mainly fostered by the PASCAL Recognizing Textual Entailment Challenge (RTE). A substantial number of papers on these topics have been published in major conferences and workshops in the last couple of years.

The primary goals of this tutorial are to review the framework of applied Textual Entailment and motivate it as a generic paradigm for natural language semantics. We will present some of the key computational approaches proposed and some of the obstacles identified by the research community in this area, as a way to promote further research. The tutorial will thus be useful for many of the senior and junior researchers that have prior or new interest in this area, providing a concise overview of recent perspectives and research results.

Tutorial Outline

1. Motivation and Task Definition
   - Textual Entailment as a generic (application independent) semantic inference test.
2. A Skeletal review of Textual Entailment Systems
   - A survey of existing approaches: a unified perspective.
3. Knowledge Acquisition Methods
4. Applications of Textual Entailment
   - Proposing ways to use generic entailment models in specific applications.
5. A Textual Entailment view of Semantics
   - Textual Entailment as a vehicle for the study of old and new semantic phenomena.