ACL-IJCNLP 2009

Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP

Tutorial Abstracts

2 August 2009
Suntec, Singapore

Production and Manufacturing by
World Scientific Publishing Co Pte Ltd
5 Toh Tuck Link
Singapore 596224

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 Co-Chairs:

Diana McCarthy, University of Sussex, UK
Chengqing Zong, Institute of Automation, Chinese Academy of Sciences (CASIA), China

Table of Contents

Tutorial Program

Sunday, August 2, 2009

Morning	*Fundamentals of Chinese Language Processing*
8:30–12:00	Chu-Ren Huang and Qin Lu
	Topics in Statistical Machine Translation
	Kevin Knight and Philipp Koehn
	Semantic Role Labeling: Past, Present and Future
	Lluís Màrquez
Afternoon	*Computational Modeling of Human Language Acquisition*
2:00–5:30	Afra Alishahi
	Learning to Rank
	Hang Li
	State-of-the-art NLP Approaches to Coreference Resolution: Theory and Practical Recipes
	Simone Paolo Ponzetto and Massimo Poesio

Fundamentals of Chinese Language Processing

Chu-Ren Huang
Dept. of Chinese and Bilingual Studies
Hong Kong polytechnic University
Churen.huang@inet.polyu.edu.hk

Qin Lu
Department of Computing
Hong Kong Polytechnic University
csluqin@comp.polyu.edu.hk

1 Introduction

This tutorial gives an introduction to the fundamentals of Chinese language processing for text processing. Today, more and more Chinese information are available in electronic form and over the internet. Computer processing of Chinese text requires the understanding of both the language itself and the technology to handle them. This tutorial is targeted for both Chinese linguists who are interested in computational linguistics and computer scientists who are interested in research on processing Chinese.

2 Content Overview

This tutorial consists of two parts. The first part overviews the grammar of the Chinese language from a language processing perspective based on naturally occurring data. The second part overviews Chinese specific processing issues and corresponding computational technologies.

The grammar introduced is a descriptive grammar of general-purpose, present-day standard Mandarin Chinese, which is fast becoming an internationally spoken language. Real examples of actual language use will be illustrated based on a data driven and corpus based approach so that its links to computational linguistic approaches for computer processing are naturally bridged in. A number of important Chinese NLP resources are also presented. On the technology side, the tutorial mainly covers Chinese word segmentation and Part-of-Speech tagging. Word segmentation problem has to deal with some Chinese language unique problems such as unknown word detection and named entity recognition which are the emphasis of this tutorial.

3 Tutorial Outline

Part 1: Highlights of Chinese Grammar for NLP
- 1.1 Preliminaries: Orthography and writing conventions
- 1.2 Basic unit of processing: word or character?
  - a. Word-forms vs. character forms
  - b. Word-senses vs. character-senses
- 1.3 Part-of-Speech: important issues in defining word classes
- 1.4 Word formation: from affixation to compounding
- 1.5 Unique constructions and challenges
  - a. Classifier-noun agreement
  - b. Separable compounds (or ionization)
  - c. 'Verbless' Constructions
- 1.6. Chinese NLP resources

Part 2: Text Processing
- 2.1 Lexical processing
  - a. Segmentation
  - b. Disambiguation
  - c. Unknown word detection
  - d. Named Entity Recognition
- 2.2 Syntactic processing
  - a. Issues in PoS tagging
  - b. Hidden Markov Models
- 2.3 NLP Applications

References

Academia Sinica Balance Corpus of Mandarin Chinese. http://www.sinica.edu.tw/SinicaCorpus/

Chao, Y. R. 1968. A Grammar of Spoken Chinese. Berkeley: University of California Press.

Huang, C.-R., K.-j. Chen and B. K. T'sou. 1996. Readings in Chinese Natural Language Processing. *Journal of Chinese Linguistics Monograph Series No. 9.* Berkeley: POLA.

T'sou, B. K. 2004. Chinese Language Processing at the Dawn of the 21st Century. In C.-R. Huang and W. Lenders. Eds. Computational Linguistics and Beyond. Pp. 189-206. Taipei: AcademiaSinica.

Miao, S.Q., Wei, Z.H. 2007, Chinese Text Information Processing Principles and Applications (In Chinese). Tsinghua University Press.

Tutorial Abstracts of ACL-IJCNLP 2009, page 1,
Suntec, Singapore, 2 August 2009.

Topics in Statistical Machine Translation

Kevin Knight
Information Sciences Institute
University of Southern California
knight@isi.edu

Philipp Koehn
School of Informatics
University of Edinburgh
pkoehn@inf.ed.ac.uk

1 Introduction

In the past, we presented tutorials called "Introduction to Statistical Machine Translation", aimed at people who know little or nothing about the field and want to get acquainted with the basic concepts. This tutorial, by contrast, goes more deeply into selected topics of intense current interest. We aim at two types of participants:

1. People who understand the basic idea of statistical machine translation and want to get a survey of hot-topic current research, in terms that they can understand.
2. People associated with statistical machine translation work, who have not had time to study the most current topics in depth.

We fill the gap between the introductory tutorials that have gone before and the detailed scientific papers presented at ACL sessions.

2 Tutorial Outline

Below is our tutorial structure. We showcase the intuitions behind the algorithms and give examples of how they work on sample data. Our selection of topics focuses on techniques that deliver proven gains in translation accuracy, and we supply empirical results from the literature.

1. QUICK REVIEW (15 minutes)
   - Phrase-based and syntax-based MT.
2. ALGORITHMS (45 minutes)
   - Efficient decoding for phrase-based and syntax-based MT (cube pruning, forward/outside costs).
   - Minimum-Bayes risk.
   - System combination.
3. SCALING TO LARGE DATA (30 minutes)
   - Phrase table pruning, storage, suffix arrays.
   - Large language models (distributed LMs, noisy LMs).
4. NEW MODELS (1 hour and 10 minutes)
   - New methods for word alignment (beyond GIZA++).
   - Factored models.
   - Maximum entropy models for rule selection and re-ordering.
   - Acquisition of syntactic translation rules.
   - Syntax-based language models and target-language dependencies.
   - Lattices for encoding source-language uncertainties.
5. LEARNING TECHNIQUES (20 minutes)
   - Discriminative training (perceptron, MIRA).

Tutorial Abstracts of ACL-IJCNLP 2009, page 2,
Suntec, Singapore, 2 August 2009.

Semantic Role Labeling: Past, Present and Future

Lluís Màrquez
TALP Research Center
Software Department
Technical University of Catalonia
lluism@lsi.upc.edu

1 Introduction

Semantic Role Labeling (SRL) consists of, given a sentence, detecting basic event structures such as "who" did "what" to "whom", "when" and "where". From a linguistic point of view, a key component of the task corresponds to identifying the semantic arguments filling the roles of the sentence predicates. Typical predicate semantic arguments include Agent, Patient, and Instrument, but semantic roles may also be found as adjuncts (e.g., Locative, Temporal, Manner, and Cause). The identification of such event frames holds potential for significant impact in many NLP applications, such as Information Extraction, Question Answering, Summarization and Machine Translation.

Recently, the compilation and manual annotation with semantic roles of several corpora has enabled the development of supervised statistical approaches to SRL, which has become a well-defined task with a substantial body of work and comparative evaluation. Significant advances in many directions have been reported over the last several years, including but not limited to: machine learning algorithms and architectures specialized for the task, feature engineering, inference to force coherent solutions, and system combinations.

However, despite all the efforts and the considerable degree of maturity of the SRL technology, the use of SRL systems in real-world applications has so far been limited and, certainly, below the initial expectations. This fact has to do with the weaknesses and limitations of current systems, which have been highlighted by many of the evaluation exercises and keep unresolved for a few years (e.g., poor generalization across corpora, low scalability and efficiency, knowledge poor features, too high complexity, absolute performance below 90%, etc.).

2 Content Overview and Outline

This tutorial has two differentiated parts. In the first one, the state-of-the-art on SRL will be overviewed, including: main techniques applied, existing systems, and lessons learned from the CoNLL and SemEval evaluation exercises. This part will include a critical review of current problems and the identification of the main challenges for the future. The second part is devoted to the lines of research oriented to overcome current limitations. This part will include an analysis of the relation between syntax and SRL, the development of joint systems for integrated syntactic-semantic analysis, generalization across corpora, and engineering of truly semantic features. See the outline below.

1. Introduction
   - Problem definition and properties
   - Importance of SRL
   - Main computational resources and systems available for SRL
2. State-of-the-art SRL systems
   - Architecture
   - Training of different components
   - Feature engineering
3. Empirical evaluation of SRL systems
   - Evaluation exercises at SemEval and CoNLL conferences
   - Main lessons learned
4. Current problems and challenges
5. Keys for future progress
   - Relation to syntax: joint learning of syntactic and semantic dependencies
   - Generalization across domains and text genres
   - Use of semantic knowledge
   - SRL systems in applications
6. Conclusions

Tutorial Abstracts of ACL-IJCNLP 2009, page 3,
Suntec, Singapore, 2 August 2009.

Computational Modeling of Human Language Acquisition

Afra Alishahi
Department of Computational Linguistics and Phonetics
Saarland University, Germany
afra@coli.uni-saarland.de

1 Introduction

The nature and amount of information needed for learning a natural language, and the underlying mechanisms involved in this process, are the subject of much debate: is it possible to learn a language from usage data only, or some sort of innate knowledge and/or bias is needed to boost the process? This is a topic of interest to (psycho)linguists who study human language acquisition, as well as computational linguists who develop the knowledge sources necessary for largescale natural language processing systems. Children are a source of inspiration for any such study of language learnability. They learn language with ease, and their acquired knowledge of language is flexible and robust.

Human language acquisition has been studied for centuries, but using computational modeling for such studies is a relatively recent trend. However, computational approaches to language learning have become increasing popular, mainly due to the advances in developing machine learning techniques, and the availability of vast collections of experimental data on child language learning and child-adult interaction. Many of the existing computational models attempt to study the complex task of learning a language under the cognitive plausibility criteria (such as memory and processing limitations that humans face), as well as to explain the developmental patterns observed in children. Such computational studies can provide insight into the plausible mechanisms involved in human language acquisition, and be a source of inspiration for developing better language models and techniques.

2 Content Overview

This tutorial will discuss the main research questions that the researchers in the field of computational language acquisition are concerned with, and will review common approaches and techniques used in developing such models. Computational modeling has been vastly applied to different domains of language acquisition, including word segmentation and phonology, morphology, syntax, semantics and discourse. However, due to time restrictions, the focus of the tutorial will be on the acquisition of word meaning, syntax, and the relationship between syntax and semantics.

The first part of the tutorial focuses on some of the fundamental issues in the study of human language acquisition, and the role of computational modeling in addressing these issues. Specifically, we discuss language modularity, i.e. the representation and acquisition of various aspects of language, and the interaction between these aspects. We also review the major arguments on language learnability and innateness. We then give a general overview of how computational modeling is used for investigating different views on each of these topics, how the theoretical assumptions are integrated into computational models, and how such models are evaluated based on the experimental observations.

In the second part of the tutorial, we will take a closer look at some of the existing models of language learning. We discuss general trends in computational modeling over the past decades, including symbolic, connectionist, and probabilistic modeling. We review a number of more influential models of the acquisition of syntax and semantics, and the link between the two. Finally, we explore some of the available tools and resources for implementing and evaluating computational models of language acquisition.

Tutorial Abstracts of ACL-IJCNLP 2009, page 4,
Suntec, Singapore, 2 August 2009.

Learning to Rank

Hang Li
Microsoft Research Asia
4F Sigma Building, No 49 Zhichun Road, Haidian, Beijing China
hangli@microsoft.com

1 Introduction

In this tutorial I will introduce 'learning to rank', a machine learning technology on constructing a model for ranking objects using training data. I will first explain the problem formulation of learning to rank, and relations between learning to rank and the other learning tasks. I will then describe learning to rank methods developed in recent years, including pointwise, pairwise, and listwise approaches. I will then give an introduction to the theoretical work on learning to rank and the applications of learning to rank. Finally, I will show some future directions of research on learning to rank. The goal of this tutorial is to give the audience a comprehensive survey to the technology and stimulate more research on the technology and application of the technology to natural language processing.

Learning to rank has been successfully applied to information retrieval and is potentially useful for natural language processing as well. In fact many NLP tasks can be formalized as ranking problems and NLP technologies may be significantly improved by using learning to rank techniques. These include question answering, summarization, and machine translation. For example, in machine translation, given a sentence in the source language, we are to translate it to a sentence in the target language. Usually there are multiple possible translations and it would be better to sort the possible translations in descending order of their likelihood and output the sorted results. Learning to rank can be employed in the task.

2 Outline

1. Introduction
2. Learning to Rank Problem
   (a) Problem Formulation
   (b) Evaluation
3. Learning to Rank Methods
   (a) Pointwise Approach
      i. McRank
   (b) Pairwise Approach
      i. Ranking SVM
      ii. RankBoost
      iii. RankNet
      iv. IR SVM
   (c) Listwise Approach
      i. ListNet
      ii. ListMLE
      iii. AdaRank
      iv. SVM Map
      v. PermuRank
      vi. SoftRank
   (d) Other Methods
4. Learning to Rank Theory
   (a) Pairwise Approach
      i. Generalization Analysis
   (b) Listwise Approach
      i. Generalization Analysis
      ii. Consistency Analysis
5. Learning to Rank Applications
   (a) Search Ranking
   (b) Collaborative Filtering
   (c) Key Phrase Extraction
   (d) Potential Applications in Natural Language Processing
6. Future Directions for Learning to Rank Research
7. Conclusion

Tutorial Abstracts of ACL-IJCNLP 2009, page 5,
Suntec, Singapore, 2 August 2009.

State-of-the-art NLP Approaches to Coreference Resolution: Theory and Practical Recipes

Simone Paolo Ponzetto
Seminar für Computerlinguistik
University of Heidelberg
ponzetto@cl.uni-heidelberg.de

Massimo Poesio
DISI
University of Trento
massimo.poesio@unitn.it

1 Introduction

The identification of different nominal phrases in a discourse as used to refer to the same (discourse) entity is essential for achieving robust natural language understanding (NLU). The importance of this task is directly amplified by the field of Natural Language Processing (NLP) currently moving towards high-level linguistic tasks requiring NLU capabilities such as e.g. recognizing textual entailment. This tutorial aims at providing the NLP community with a gentle introduction to the task of coreference resolution from both a theoretical and an application-oriented perspective. Its main purposes are: (1) to introduce a general audience of NLP researchers to the core ideas underlying state-of-the-art computational models of coreference; (2) to provide that same audience with an overview of NLP applications which can benefit from coreference information.

2 Content Overview

1. Introduction to machine learning approaches to coreference resolution. We start by focusing on machine learning based approaches developed in the seminal works from Soon et al. (2001) and Ng & Cardie (2002). We then analyze the main limitations of these approaches, i.e. their clustering of mentions from a local pairwise classification of nominal phrases in text. We finally move on to present more complex models which attempt to model coreference as a global discourse phenomenon (Yang et al., 2003; Luo et al., 2004; Daumé III & Marcu, 2005, inter alia).

2. Lexical and encyclopedic knowledge for coreference resolution. Resolving anaphors to their correct antecedents requires in many cases lexical and encyclopedic knowledge. We accordingly introduce approaches which attempt to include semantic information into the coreference models from a variety of knowledge sources, e.g. WordNet (Harabagiu et al., 2001), Wikipedia (Ponzetto & Strube, 2006) and automatically harvested patterns (Poesio et al., 2002; Markert & Nissim, 2005; Yang & Su, 2007).

3. Applications and future directions. We present an overview of NLP applications which have been shown to profit from coreference information, e.g. question answering and summarization. We conclude with remarks on future work directions. These include: a) bringing together approaches to coreference using semantic information with global discourse modeling techniques; b) exploring novel application scenarios which could potentially benefit from coreference resolution, e.g. relation extraction and extracting events and event chains from text.

References

Daumé III, H. & D. Marcu (2005). A large-scale exploration of effective global features for a joint entity detection and tracking model. In *Proc. HLT-EMNLP '05*, pp. 97–104.

Harabagiu, S. M., R. C. Bunescu & S. J. Maiorano (2001). Text and knowledge mining for coreference resolution. In *Proc. of NAACL-01*, pp. 55–62.

Luo, X., A. Ittycheriah, H. Jing, N. Kambhatla & S. Roukos (2004). A mention-synchronous coreference resolution algorithm based on the Bell Tree. In *Proc. of ACL-04*, pp. 136–143.

Markert, K. & M. Nissim (2005). Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31(3):367–401.

Ng, V. & C. Cardie (2002). Improving machine learning approaches to coreference resolution. In *Proc. of ACL-02*, pp. 104–111.

Poesio, M., T. Ishikawa, S. Schulte im Walde & R. Vieira (2002). Acquiring lexical knowledge for anaphora resolution. In *Proc. of LREC '02*, pp. 1220–1225.

Ponzetto, S. P. & M. Strube (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proc. of HLT-NAACL-06*, pp. 192–199.

Soon, W. M., H. T. Ng & D. C. Y. Lim (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Yang, X. & J. Su (2007). Coreference resolution using semantic relatedness information from automatically discovered patterns. In *Proc. of ACL-07*, pp. 528–535.

Yang, X., G. Zhou, J. Su & C. L. Tan (2003). Coreference resolution using competition learning approach. In *Proc. of ACL-03*, pp. 176–183.

Tutorial Abstracts of ACL-IJCNLP 2009, page 6,
Suntec, Singapore, 2 August 2009.

Author Index