The 2013 Conference of the
North American Chapter of the
Association for Computational Linguistics:
Human Language Technologies

Conference Handbook

June 9 - 15, 2012
Atlanta, Georgia

This year with

the collocated
Second Joint Conference on
Lexical and Computational
Semantics
June 13–14, 2013

Message from the General Chair

Welcome everyone!

It is my pleasure to welcome you all to Atlanta, Georgia, for the 2013 NAACL Human Language Technologies conference. This is a great opportunity to reconnect with old friends and make new acquaintances, learn the latest in your own field and become curious about new areas, and also to experience Atlanta's warm southern hospitality. That hospitality starts with Priscilla Rasmussen! Priscilla thinks about everything that we all take for granted: the registration that just took place, the rooms in which we sit, the refreshments that keep us energized, and the social events that make this conference so fun, and many other details that you would miss if they weren't there. Please introduce yourself and say hi. Priscilla is the backbone of the NAACL organization. Thank you!

This conference started a year ago, when Hal Daumé III and Katrin Kirchhoff graciously agreed to be program co-chairs. It is no exaggeration to say how much their dedication has shaped this conference and how grateful I am for their initiative and hard work. Thank you Hal and Katrin, especially for all the fun discussion that made the work light and the year go by fast! This conference could not have happened with you.

Thanks go to the entire organizing committee. As I am writing this to be included in the proceedings, I am grateful for the fantastic detailed and proactive work by Colin Cherry and Matt Post, the publications chairs. The tutorials chairs, Katrin Erk and Jimmy Lin, selected, and solicited, 6 tutorials to present in depth material on some of the diverse topics represented in our community. Chris Dyer and Derrick Higgins considered which projects shine best when shown as a demonstration. The workshops chairs for NAACL, Sujith Ravi and Luke Zettlemoyer, worked jointly with ACL and EMNLP to select the workshops to be held at NAACL. They also worked with ICML 2013 to co-host workshops that bridge the two communities, in addition to the Joint NAACL/ICML symposium.

Posters from the student research workshop are part of the poster and demonstrations session on Monday night. This is a great opportunity for the students to be recognized in the community and to benefit from lively discussion of their presentations (attendees take note!) Annie Louis and Richard Socher are the student research workshop chairs, and Julia Hockenmaier and Eric Ringger generously share their wisdom as the faculty advisors. The student research workshop itself will be held on the first day of workshops. There are so many people who contribute their time to the behind-the-scenes organization of the conference, without which the conference cannot take place. Asking for money is probably not a natural fit for anyone, but Chris Brew worked on local sponsorship, and Dan Bikel and Patrick Pantel worked

to obtain sponsorship across the ACL conferences this year - thank you! Jacob Eisenstein had the more fun role of distributing money as the student volunteer coordinator, and we thank all of the student volunteers who will be helping to run a smooth conference. Kristy Boyer kept the communication "short and tweet" using a variety of social media (and old-fashioned media too). An important part of the behind-the-scenes efforts that enable a conference like NAACL to come together are the sponsors. We thank all of the sponsors for the contributions to the conference , both for the general funding made available as well as the specific programs that are funded through sponsorship. You can read more about these sponsors in our conference handbook.

This year there are several initiatives, and if successful, we hope they'll be part of NAACL conferences in the future. One is to make the proceedings available prior to the conference; we hope you will benefit from the extra time to read the papers beforehand. Another is for tutorials and all oral presentations to be recorded on video and made available post-conference. We are also delighted to host presentations, in both oral and poster formats, from the new Transactions of the ACL journal, to enhance the impact these will already have as journal publications. Finally, Matt Post is creating a new digital form of conference handbook to go with our digital age; thanks also go to Alex Clemmer who has prepared the paper copy that you may be reading right now. We hope you use the #NAACL2013 tag when you are tweeting about the conference or papers at the conference; together, we'll be creating a new social media corpus to explore.

Once again, we are pleased to be co-located with *SEM conference, and the SemEval workshop. We are lucky to have ICML 2013 organized so close in time and place. Several researchers who span the two communities have reconvened the Joint NAACL/ICML symposium on June 15, 2013. In addition, two workshops that address areas of interest to both NAACL and ICML members have been organized on June 16th, as part of the ICML conference.

NAACL 2013 has given me a great appreciation for the volunteering that is part of our culture. Besides the organizing committee itself, we are guided by the NAACL executive board, who think about questions with a multi-year perspective. I also want to recognize the members who first initiated and now maintain the ACL Anthology, where all of our published work will be available to all in perpetuity, a fabulous contribution and one that distinguishes our academic community.

Have a fun conference!

Lucy Vanderwende, Microsoft Research
NAACL HLT 2013 General Chair

Message from the Program Committee Co-Chairs

Welcome to NAACL HLT 2013 in Atlanta, Georgia. We have an exciting program consisting of six tutorials, 24 sessions of talks (both for long and short papers), an insane poster madness session that includes posters from the newly revamped student research workshop, ten workshops and two additional cross-pollination workshops held jointly with ICML (occurring immediately after NAACL HLT, just one block away). There are a few innovations in the conference this year, the most noticeable of which is the twitter channel #naacl2013 and the fact that we are the first conference to host papers published in the Transactions of the ACL journal – there are six such papers in our program, marked as [TACL]. We are very excited about our two invited talks, one on Monday morning and one Wednesday morning. The first is by Gina Kuperberg, who will talk about "Predicting Meaning: What the Brain tells us about the Architecture of Language Comprehension." The second presenter is our own Kathleen KcKeown, who will talk about "Natural Language Applications from Fact to Fiction."

The morning session on Tuesday includes the presentation of best paper awards to two worthy recipients. The award for Best Short Paper goes to Marta Recasens, Marie-Catherine de Marneffe and Christopher Potts for their paper "The Life and Death of Discourse Entities: Identifying Singleton Mentions" The award for Best Student Paper goes to the long paper "Automatic Generation of English Respellings" by Bradley Hauer and Greg Kondrak. We gratefully acknowledge IBM's support for the Student Best Paper Award. Finally, many thanks to the Best Paper Committee for selecting these excellent papers!

The complete program includes 95 long papers (of which six represent presentations from the journal Transactions of the ACL, a first for any ACL conference!) and 51 short papers. We are excited that the conference is able to present such a dynamic array of papers, and would like to thank the authors for their great work. We worked hard to keep the conference to three parallel sessions at any one time to hopefully maximize a participant's ability to see everything she wants! This represents an acceptance rate of 30% for long papers and 37% for short papers. More details about the distribution across areas and other statistics will be made available in the NAACL HLT Program Chair report on the ACL wiki: `aclweb.org/adminwiki/index.php?title=Reports`

The review process for the conference was double-blind, and included an author response period for clarifying reviewers' questions. We were very pleased to have the assistance of 350 reviewers, each of whom reviewed an average of 3.7 papers, in deciding the program. We are especially thankful for the reviewers who spent time reading the author responses and engaging other reviewers in the discussion board. Assigning reviewers would not have been possible without the hard work of Mark Dredze and his miracle assignment scripts. Furthermore, constructing the program would not

have been possible without 22 excellent area chairs forming the Senior Program Committee: Eugene Agichtein, Srinivas Bangalore, David Bean, Phil Blunsom, Jordan Boyd-Graber, Marine Carpuat, Joyce Chai, Vera Demberg, Bill Dolan, Doug Downey, Mark Dredze, Markus Dreyer, Sanda Harabagiu, James Henderson, Guy Lapalme, Alon Lavie, Percy Liang, Johanna Moore, Ani Nenkova, Joakim Nivre, Bo Pang, Zak Shafran, David Traum, Peter Turney, and Theresa Wilson. Area chairs were responsible for managing paper assignments, collating reviewer responses, handling papers for other area chairs or program chairs who had conflicts of interest, making recommendations for paper acceptance or rejection, and nominating best papers from their areas. We are very grateful for the time and energy that they have put into the program.

There are a number of other people that we interacted with who deserve a hearty thanks for the success of the program. Rich Gerber and the START team at Softconf have been invaluable for helping us with the mechanics of the reviewing process. Matt Post and Colin Cherry, as publications cochairs, have been very helpful in assembling the final program and coordinating the publications of the workshop proceedings. There are several crucial parts of the overall program that were the responsibility of various contributors, including Annie Louis, Richard Socher, Julia Hockenmaier and Eric Ringger (Student Research Workshop chairs, who did an amazing job revamping the SRW); Jimmy Lin and Katrin Erk (Tutorial Chairs); Luke Zettlemoyer and Sujith Ravi (Workshop Chairs); Chris Dyer and Derrick Higgins (Demo Chairs); Jacob Eisenstein (Student Volunteer Coordinator); Chris Brew (Local Sponsorship Chair); Patrick Pantel and Dan Bikel (Sponsorship Chairs); and the new-founded Publicity chair who handled #naacl2013 tweeting among other things, Kristy Boyer.

We would also like to thank Chris Callison-Burch and the NAACL Executive Board for guidance during the process. Michael Collins was amazingly helpful in getting the inaugural TACL papers into the NAACL HLT conference. Priscilla Rasmussen deserves, as always, special mention and warmest thanks as the local arrangements chair and general business manager. Priscilla is amazing and everyone who sees her at the conference should thank her.

Finally, we would like to thank our General Chair, Lucy Vanderwende, for both her trust and guidance during this process. She helped turn the less-than-wonderful parts of this job to roses, and her ability to organize an incredibly complex event is awe inspiring. None of this would have happened without her.

We hope that you enjoy the conference!

Hal Daumé III, University of Maryland
Katrin Kirchhoff, University of Washington

NAACL HLT 2013 Organizing Committee

General Conference Chair
Lucy Vanderwende, Microsoft Research

Program Committee Chairs
Hal Daumé III, University of Maryland
Katrin Kirchhoff, University of Washington

Local Arrangements
Priscilla Rassmussen

Workshop Chairs
Luke Zettlemoyer, University of Washington
Sujith Ravi, Google

Tutorial Chairs
Jimmy Lin, University of Maryland
Katrin Erk, University of Texas at Austin

Student Research Workshop
Chairs:
    Annie Louis, University of Pennsylvania
    Richard Socher, Stanford University
Faculty Advisors:
    Julia Hockenmaier, University of Illinois at Urbana-Champaign
    Eric Ringger, Brigham Young University

Student Volunteer Coordinator
Jacob Eisenstein, School of Interactive Computing, Georgia Tech

Demonstrations Chairs
Chris Dyer, Carnegie Mellon University
Derrick Higgins, Educational Testing Service

Local Sponsorship Chair
Chris Brew, Educational Testing Service

NAACL Sponsorship Chairs
Patrick Pantel, Microsoft Research
Dan Bikel, Google

Publications Chairs
Matt Post, Johns Hopkins University
Colin Cherry, National Research Council, Canada

Publicity Chair
Kristy Boyer, North Carolina State University

Area Chairs
Phonology and Morphology, Word Segmentation
Markus Dreyer (SDL Language Weaver)

Syntax, Tagging, Chunking and Parsing
Joakim Nivre (Uppsala University)

James Henderson (Université de Genève)

Semantics
Percy Liang (Stanford University)
Peter Turney (National Research Council of Canada)

Multimodal NLP
Srinivas Bangalore (AT&T)

Discourse, Dialogue, Pragmatics
David Traum (Institute for Creative Technologies)
Joyce Chai (Michigan State University)

Linguistic Aspects of CL
Vera Demberg (Saarland University)

Summarization
Guy Lapalme (Université de Montréal)

Generation
Johanna Moore (University of Edinburgh)

ML for Language Processing
Phil Blunsom (University of Oxford)
Mark Dredze (Johns Hopkins University)

Machine Translation
Alon Lavie (Carnegie Mellon University)
Marine Carpuat (National Research Council of Canada)

Information Retrieval and QA
Eugene Agichtein (Emory University)

Information Extraction
Doug Downey (Northwestern University)
Sanda Harabagiu (University of Texas at Dallas)

Spoken Language Processing
Zak Shafran (Oregon Health and Science University)

Sentiment Analysis and Opinion Mining
Bo Pang (Cornell University)
Theresa Wilson (Johns Hopkins University)

NLP-enabled Technology
David Bean (TDW)

Document Categorization and Topic Clustering
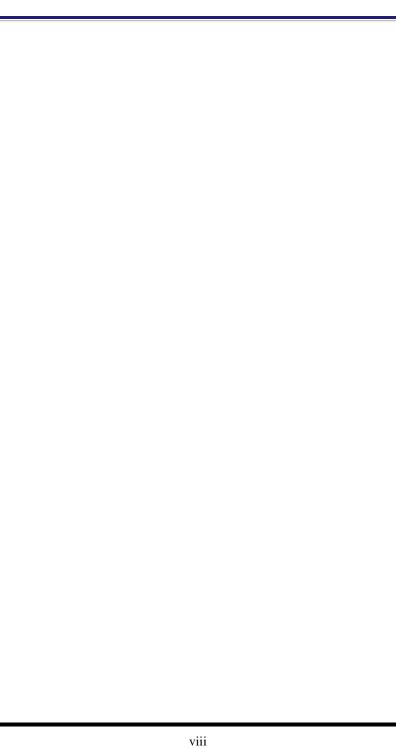Jordan Boyd-Graber (University of Maryland)

Social Media Analysis and Processing
Bill Dolan (Microsoft Research)

Language Resources and Evaluation Methods
Ani Nenkova (University of Pennsylvania)

Contents

LOCAL INFO HERE

1

Overview

Sunday, June 9, 2013

           Morning Session

9:00–12:30   *Deep Learning for NLP (without Magic)*
           Richard Socher and Christopher D. Manning

9:00–12:30   *Discourse Processing*
           Manfred Stede

9:00–12:30   *NLP for uncertain data at scale*
           Sameep Mehta and L V Subramaniam

           Afternoon Session

2:00–5:30   *Semantic Role Labeling*
           Martha Palmer, Ivan Titov and Shumin Wu

2:00–5:30   *Spectral Learning Algorithms for Natural Language Processing*
           Shay Cohen, Michael Collins, Dean Foster, Karl Stratos and Lyle Ungar

2:00–5:30   *Morphological, Syntactical and Semantic Knowledge in Statistical Machine Translation*
           Marta Ruiz Costa-jussà and Chris Quirk

Tutorial 1

Deep Learning for NLP (without Magic)

Richard Socher and Christopher D. Manning

Sunday, June 9, 2013, 9:00am – 12:30pm

International B,C

Abstract: Machine learning is everywhere in today's NLP, but by and large machine learning amounts to numerical optimization of weights for human designed representations and features. The goal of deep learning is to explore how computers can take advantage of data to develop features and representations appropriate for complex interpretation tasks. This tutorial aims to cover the basic motivation, ideas, models and learning algorithms in deep learning for natural language processing. Recently, these methods have been shown to perform very well on various NLP tasks such as language modeling, POS tagging, named entity recognition, sentiment analysis and paraphrase detection, among others. The most attractive quality of these techniques is that they can perform well without any external hand-designed resources or time-intensive feature engineering. Despite these advantages, many researchers in NLP are not familiar with these methods. Our focus is on insight and understanding, using graphical illustrations and simple, intuitive derivations. The goal of the tutorial is to make the inner workings of these techniques transparent, intuitive and their results interpretable, rather than black boxes labeled "magic here". The first part of the tutorial presents the basics of neural networks, neural word vectors, several simple models based on local windows and the math and algorithms of training via backpropagation. In this section applications include language modeling and POS tagging. In the second section we present recursive neural networks which can learn structured tree outputs as well as vector representations for phrases and sentences. We cover both equations as well as applications. We show how training can be achieved by a modified version of the backpropagation algorithm introduced before. These modifications allow the algorithm to work on tree structures. Applications include sentiment analysis and paraphrase detection. We also draw connections to recent work in semantic compositionality in vector spaces. The principle goal, again, is to make these methods appear intuitive and interpretable rather than mathematically confusing. By this point in the tutorial, the audience members should have a clear understanding of how to build a deep learning system for word-, sentence- and document-level tasks. The last part of the tutorial gives a general overview of the different applications of deep learning in NLP, including bag of

---

Richard Socher is a PhD student at Stanford working with Chris Manning and Andrew Ng. His research interests are machine learning for NLP and vision. He is interested in developing new models that learn useful features, capture compositional and hierarchical structure in multiple modalities and perform well across different tasks. He was awarded the 2011 Yahoo! Key Scientific Challenges Award, the Distinguished Application Paper Award at ICML 2011 and a Microsoft Research PhD Fellowship in 2012.

Christopher Manning is an Associate Professor of Computer Science and Linguistics at Stanford University (PhD, Stanford, 1995). Manning has coauthored leading textbooks on statistical approaches to NLP (Manning and Schuetze 1999) and information retrieval (Manning et al. 2008). His recent work concentrates on machine learning and natural language processing, including applications such as statistical parsing and text understanding, joint probabilistic inference, clustering, and deep learning over text and images.

words models. We will provide a discussion of NLP-oriented issues in modeling, interpretation, representational power, and optimization.

Tutorial 2

Discourse Processing

Manfred Stede (University of Potsdam)

Sunday, June 9, 2013, 9:00am – 12:30pm

International D

Abstract: The observation that discourse is more than a mere sequence of utterances or sentences amounts to a truism. But what follows from this? In what way does the "value added" arise when segments of discourse are juxtaposed - how does hierarchical structure originate from a linearized discourse?

While many discourse phenomena apply to dialogue and monologue alike, this tutorial will center its attention on monologue written text. The perspective taken is that of practical language processing: We study methods for automatically deriving discourse information from text, and point to aspects of their implementation. The emphasis is on breadth rather than depth, so that the attendees will get an overview of the central tasks of discourse processing, with pointers to the literature for studying the individual problems in more depth. Much of the tutorial will follow the line of the recent book M. Stede: Discourse Processing. Morgan & Claypool 2011.

Specifically, we will study the most important ways of ascribing structure to discourse. This is, first, a breakdown into functional units that are characteristic for the genre of the text. A news message, for example, is conventionally structured in a different way than a scientific paper is. For grasping this level of structure, the patterns that are characteristic for the specific genre need to be modeled.

Second, an ongoing text, unless it is very short, will cover different topics and address them in a sensible linear order. This is largely independent of genre, and since the notion of topic is relatively vague, it is harder to describe and sometimes difficult to identify. The common approach is to track the distribution of content words across the text, but in addition, overt signals for topic switches can be exploited.

Third, the identification of coreference links is a central aspect of discourse processing, and has received much attention in computational linguistics. We will survey the corpus-based methods

---

Manfred Stede, University of Potsdam. After completing his dissertation on the role of lexical semantics in multilingual text generation, Manfred Stede shifted his research focus towards problems of discourse structure and its role in various applications of text understanding. For discourse structure, his work centered on coherence relations and associated structural descriptions of text, and on the linguistic signals of such relations, especially connectives. From the early 2000s on, he developed the Potsdam Commentary Corpus as an example of (German) texts analyzed simultaneously on multiple levels, including sentential syntax, coreference, and rhetorical structure; in parallel, the technical infrastructure of a database for querying and visualizing multi-layer corpora was developed. In recent years, more analysis levels have been added to the corpus (e.g., content zones, connectives and their arguments). As for applications, Manfred worked on text summarization and various tasks of information extraction; more recently, his focus has been on issues of subjectivity and sentiment analysis.

that have dominated the field in recent years, and then look at the ramifications that the set of all coreference links in a text has for its structure.

Fourth, we investigate the structure resulting from establishing coherence relations (e.g., Cause, Contrast) among adjacent text segments. The term "discourse parsing" is often used for the task of identifying such relations (by exploiting more or less explicit linguistic signals) and building tree structures that reflect the semantic or pragmatic scaffolding of a (portion of) text.

Thus emerges a picture of a text as a series of different, yet related, layers of analysis. The final part of the tutorial addresses the issue of inter-connections between these levels. As a tool for accessing such multi-layered text corpora, we will see how the (open-source) ANNIS2 database allows for querying the data across different layers, and for visualizing different structural layers in appropriate ways.

Tutorial 3

Towards Reliability-Aware Entity Analytics and Integration for Noisy Text at Scale

Sameep Mehta and L. Venkata Subramaniam

Sunday, June 9, 2013, 9:00am – 12:30pm

International E,F

Abstract: Due to easy to use apps (Facebook, Twitter, etc.), higher Internet connectivity and always on facility allowed by smart phones, the key characteristics of raw data are changing. This new data can be characterized by 4V's - Volume, Velocity, Variety and Veracity. For example during a Football match, some people will Tweet about goals, penalties, etc., while others may write longer blogs and further there will be match reports filed in trusted online news media after the match. Although the sources may be varied, the data describes and provides multiple evidences for the same event. Such multiple evidences should be used to strengthen the belief in the underlying physical event as the individual data points may have inherent uncertainty. The uncertainty can arise from inconsistent, incomplete and ambiguous reports. The uncertainty is also because the trust levels of the different sources vary and affect the overall reliability. We will summarize various efforts to perform reliability aware entity integration.

The other problem in text analysis in such setting is posed by presence of noise in the text. Since the text is produced in several informal settings such as email, blogs, tweet, SMS, chat and is inherently noisy and has several veracity issues. For example, missing punctuation and the use of non-standard words can often hinder standard natural language processing techniques such as part-of-speech tagging and parsing. Further downstream applications such as entity extraction, entity resolution and entity completion have to explicitly handle noise in order to return useful results. Often, depending on the application, noise can be modeled and it may be possible to develop specific strategies to immunize the system from the effects of noise and improve performance. Also the aspect of reliability is key as a lot of this data is ambiguous, incomplete, conflicting, untrustworthy and deceptive. The key goals of this tutorial are:

Sameep Mehta is researcher in Information Management Group at IBM Research India. He received his M.S. and Ph.D. from The Ohio State University, USA in 2006. He also holds an Adjunct Faculty position at the International Institute of Information Technology, New Delhi. Sameep regularly advises MS and PhD students at University of Delhi and IIT Delhi. He regularly delivers Tutorials at COMAD (2009, 2010 and 2011). His current research interests include Data Mining, Business Analytics, Service Science, Text Mining, and Workforce Optimization.
L Venkata Subramaniam manages the information management analytics and solutions group at IBM Research India. He received his PhD from IIT Delhi in 1999. His research focuses on unstructured information management, statistical natural language processing, noisy text analytics, text and data mining, information theory, speech and image processing. He often teaches and guides student thesis at IIT Delhi on these topics. His tutorial titled Noisy Text Analytics was the second largest at NAACL-HLT 2010. He co founded the AND (Analytics for Noisy Unstructured Text Data) workshop series and also co-chaired the first four workshops, 2007-2010. He was guest co-editor of two special issues on Noisy Text Analytics in the International Journal of Document Analysis and Recognition in 2007 and 2009.

- Draw the attention of researchers towards methods for doing entity analytics and integration on data with 4V characteristics.

- Differentiate between noise and uncertainty in such data.

- Provide an in-depth discussion on handling noise in NLP based methods.

- Finally, handling uncertainty through information fusion and integration.

This tutorial builds on two earlier tutorials — NAACL 2010 tutorial on Noisy Text and CO-MAD 2012 tutorial on Reliability Aware Data Fusion. In parallel the authors are also hosting a workshop on related topic "Reliability Aware Data Fusion" at SIAM Data Mining Conference, 2013.

Tutorial 4

Semantic Role Labeling

Martha Palmer (University of Colorado), Ivan Titov (Saarland University), and Shumin Wu (University of Colorado)

Sunday, June 9, 2013, 2:00pm – 5:30pm

International B,C

Abstract: This tutorial will describe semantic role labeling, the assignment of semantic roles to eventuality participants in an attempt to approximate a semantic representation of an utterance. The linguistic background and motivation for the definition of semantic roles will be presented, as well as the basic approach to semantic role annotation of large amounts of corpora. Recent extensions to this approach that encompass light verb constructions and predicative adjectives will be included, with reference to their impact on English, Arabic, Hindi and Chinese. Current proposed extensions such as Abstract Meaning Representations and richer event representations will also be touched on.

Details of machine learning approaches will be provided, beginning with fully supervised approaches that use the annotated corpora as training material. The importance of syntactic parse information and the contributions of different feature choices, including tree kernels, will be discussed, as well as the advantages and disadvantages of particular machine learning algorithms and approaches such as joint inference. Appropriate considerations for evaluation will be presented as well as successful uses of semantic role labeling in NLP applications.

Martha Palmer is a Professor of Linguistics and Computer Science, and a Fellow of the Institute of Cognitive Science at the University of Colorado. Her current research is aimed at building domain-independent and language independent techniques for semantic interpretation based on linguistically annotated data, such as Proposition Banks. She has been the PI on NSF, NIH and DARPA projects for linguistic annotation (syntax, semantics and pragmatics) of English, Chinese, Korean, Arabic and Hindi. She has been a member of the Advisory Committee for the DARPA TIDES program, Chair of SIGLEX, Chair of SIGHAN, a past President of the Association for Computational Linguistics, and is a Co-Editor of JNLE and of LiLT and is on the CL Editorial Board. She received her Ph.D. in Artificial Intelligence from the University of Edinburgh in 1985.

Ivan Titov joined the Saarland University as a junior faculty and head of a research group in November 2009, following a postdoc at the University of Illinois at Urbana-Champaign. He received his Ph.D. in Computer Science from the University of Geneva in 2008 and his master's degree in Applied Mathematics and Informatics from the St. Petersburg State Polytechnic University (Russia) in 2003. His research interests are in statistical natural language processing (models of syntax, semantics and sentiment) and machine learning (structured prediction methods, latent variable models, Bayesian methods).

Shumin Wu is a Computer Science PhD student (advised by Dr. Martha Palmer) at the University of Colorado. His current research is aimed at developing and applying semantic mapping (aligning and jointly inferring predicate-argument structures between languages) to Chinese dropped-pronoun recovery/alignment, automatic verb class induction, and other applications relevant to machine translation.

We will also cover techniques for exploiting unlabeled corpora and transferring models across languages. These include methods, which project annotations across languages using parallel data, induce representations solely from unlabeled corpora (unsupervised methods) or exploit a combination of a small amount of human annotation and a large unlabeled corpus (semi-supervised techniques). We will discuss methods based on different machine learning paradigms, including generative Bayesian models, graph-based algorithms and bootstrapping style techniques.

Tutorial 5

Towards Reliability-Aware Entity Analytics and Integration for Noisy Text at Scale

Shay Cohen (Columbia University), Michael Collins (Columbia University), Dean P. Foster (University of Pennsylvania), Karl Stratos (Columbia University), and Lyle Ungar (University of Pennsylvania)

Sunday, June 9, 2013, 2:00am – 5:30pm

International B,C

Abstract: Recent work in machine learning and NLP has developed spectral algorithms for many learning tasks involving latent variables. Spectral algorithms rely on singular value decomposition as a basic operation, usually followed by some simple estimation method based on the method of moments. From a theoretical point of view, these methods are appealing in that they offer consistent estimators (and PAC-style guarantees of sample complexity) for several important latent-variable models. This is in contrast to the EM algorithm, which is an extremely successful approach, but which only has guarantees of reaching a local maximum of the likelihood function.

From a practical point of view, the methods (unlike EM) have no need for careful initialization, and have recently been shown to be highly efficient (as one example, in work under submission by the authors on learning of latent-variable PCFGs, a spectral algorithm performs at identical accuracy to EM, but is around 20 times faster).

In this tutorial we will aim to give a broad overview of spectral methods, describing theoretical guarantees, as well as practical issues. We will start by covering the basics of singular value decomposition and describe efficient methods for doing singular value decomposition. The SVD operation is at the core of most spectral algorithms that have been developed.

We will then continue to cover canonical correlation analysis (CCA). CCA is an early method from statistics for dimensionality reduction. With CCA, two or more views of the data are created, and they are all projected into a lower dimensional space which maximizes the correlation between the views. We will review the basic algorithms underlying CCA, give some formal results giving guarantees for latent-variable models and also describe how they have been applied recently to learning lexical representations from large quantities of unlabeled data. This idea of learning lexical representations can be extended further, where unlabeled data is used to learn underlying representations which are subsequently used as additional information for supervised training.

We will also cover how spectral algorithms can be used for structured prediction problems with sequences and parse trees. A striking recent result by Hsu, Kakade and Zhang (2009) shows that HMMs can be learned efficiently using a spectral algorithm. HMMs are widely used in NLP and speech, and previous algorithms (typically based on EM) were guaranteed to only reach a local maximum of the likelihood function, so this is a crucial result. We will review the basic mechanics of the HMM learning algorithm, describe its formal guarantees, and also cover practical issues.

Last, we will cover work about spectral algorithms in the context of natural language parsing. We will show how spectral algorithms can be used to estimate the parameter models of latent-variable PCFGs, a model which serves as the base for state-of-the-art parsing models such as the one of Petrov et al. (2007). We will show what are the practical steps that are needed to be taken in order to make spectral algorithms for L-PCFGs (or other models in general) practical and comparable to state of the art.

Tutorial 6

Morphological, Syntactical and Semantic Knowledge in Statistical Machine Translation

Marta R. Costa-jussà (Institute for Infocomm Research) and Chris Quirk (Microsoft Research)

Sunday, June 9, 2013, 2:00am – 5:30pm

International D

Abstract:

This tutorial focuses on how morphology, syntax and semantics may be introduced into a standard phrase-based statistical machine translation system with techniques such as machine learning, parsing and word sense disambiguation, among others.

Regarding the phrase-based system, we will describe only the key theory behind it. The main challenges of this approach are that the output contains unknown words, wrong word orders and non-adequate translated words. To solve these challenges, recent research enhances the standard system using morphology, syntax and semantics.

Morphologically-rich languages have many different surface forms, even though the stem of a word may be the same. This leads to rapid vocabulary growth, as various prefixes and suffixes can combine with stems in a large number of possible combinations. Language model probability estimation is less robust because many more word forms occur rarely in the data. This morphologically-induced sparsity can be reduced by incorporating morphological information into the SMT system. We will describe the three most common solutions to face morphology: preprocessing the data so that the input language more closely resembles the output language; using additional language models that introduce morphological information; and post-processing the output to add proper inflections.

Marta R. Costa-jussà, Institute for Infocomm Research (I2R), is a Telecommunication's Engineer by the Universitat Politècnica de Catalunya (UPC, Barcelona) and she received her PhD from the UPC in 2008. Her research experience is mainly in Automatic Speech Recognition, Machine Translation and Information Retrieval. She has worked at LIMSI-CNRS (Paris), Barcelona Media Innovation Center (Barcelona) and the Universidade de Sao Paulo (São Paulo). Since December 2012 she is working at Institute for Infocomm Research (Singapore) implementing the IMTraP project ("Integration of Machine Translation Paradigms") on Hybrid Machine Translation, funded by the European Marie Curie International Outgoing European Fellowship program. She is currently organizing the ACL Workshop HyTRA 2013 and she will be teaching a summer school course on hybrid machine translation at ESSLLI 2013.
Chris Quirk, Microsoft Research. After studying Computer Science and Mathematics at Carnegie Mellon University, Chris joined Microsoft in 2000 to work on the Intentional Programming project, an extensible compiler and development framework. He moved to the Natural Language Processing group in 2001, where his research has mostly focused on statistical machine translation powering Microsoft Translator, especially on several generations of a syntax directed translation system that powers over half of the translation systems. He is also interested in semantic parsing, paraphrase methods, and very practical problems such as spelling correction and transliteration.

Syntax differences between the source and target language may lead to significant differences in the relative word order of translated words. Standard phrase-based SMT systems surmount reordering/syntactic challenges by learning from data. Most approaches model reordering inside translation units and using statistical methodologies, which limits the performance in language pairs with different grammatical structures. We will briefly introduce some recent advances in SMT that use modeling approaches based on principles more powerful flat phrases and better suited to the hierarchical structures of language: SMT decoding with stochastic synchronous context free grammars and syntax-driven translation models.

Finally, semantics are not directly included in the SMT core algorithm, which means that challenges such as polysemy or synonymy are either learned directly from data or they are incorrectly translated. We will focus on recent attempts to introduce semantics into statistical-based systems by using source context information.

The course material will be suitable both for attendees with limited knowledge of the field, and for researchers already familiar with SMT who wish to learn about modern tendencies in hybrid SMT. The mathematical content of the course include probability and simple machine learning, so reasonable knowledge of statistics and mathematics is required. There will be a small amount of linguistics and ideas from natural language processing.