Typ-NLP Keynote Speakers

Emily Bender

Title: The Grammar Matrix and AGGREGATION: Knowledge Rich NLP for Endangered and Low Resource Languages

In this talk, I will describe how the Grammar Matrix project builds on typological research to facilitate the development of precision implemented grammars for diverse languages. The Grammar Matrix combines the breadth of typological analysis with the depth of formal syntactic modeling, something which is only possible with the support of computational methods. I will also discuss the work of the AGGREGATION project which seeks to combine the linguistic knowledge encoded in interlinear glossed text (IGT) with that stored in the Grammar Matrix and its customization system to automatically produce grammars for endangered and other low-resource languages.

Short Bio

Emily M. Bender's primary research interests are in multilingual grammar engineering, the study of variation, both within and across languages, and the relationship between linguistics and computational linguistics. She is the LSA's delegate to the ACL. Her 2013 book Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax aims to present linguistic concepts in a manner accessible to NLP practitioners.

Jason Eisner

Title: Improving Cross-Linguistic Robustness by Training on Synthetic Languages

A "multilingual" or "polyglot" NLP system shares some parameters across languages. Its performance on a low-resource or zero-resource target language can benefit from training on similar source languages. Can we hope to obtain more relevant source languages by concocting new ones? We systematically permute the word order in real source treebanks -- what would English look like if it were SOV? -- to obtain new supervised training data that can, in fact, be useful for zero-shot syntactic analysis. This can be done deliberately to match a particular target language, or randomly to increase the diversity of the source language pool. Several practical notions of typological features will come into play, based on properties of a corpus's part-of-speech sequences or dependency trees.
This is joint work with Dingquan Wang, based on papers in TACL (2016, 2017, 2018) and EMNLP (2018).

Short Bio

Jason Eisner is Professor of Computer Science at Johns Hopkins University. He is a Fellow of the Association for Computational Linguistics (and an action editor of its TACL journal). At Johns Hopkins, he is also affiliated with the Center for Language and Speech Processing, the Machine Learning Group, the Cognitive Science Department, and the national Center of Excellence in Human Language Technology. His goal is to develop the probabilistic modeling, inference, and learning techniques needed for a unified model of all kinds of linguistic structure. His 125+ papers have presented various algorithms for parsing, machine translation, and weighted finite-state machines; formalizations, algorithms, theorems, and empirical results in computational phonology; and unsupervised or semi-supervised learning methods for syntax, morphology, and word-sense disambiguation.

Balthasar Bickel

Title: Cross-linguistic corpora shed light on the drivers of language dynamics

A fundamental property of human language is relentless change and diversification, but what is the nature of this dynamics? Is it a random process, subject solely to historical contingency? Or is it constrained by mechanisms of how the human brain processes and learns languages? A key challenge to answering these questions is that the brain is itself a dynamic and plastic system so that many mechanisms adapt to rather than rein in the languages they are confronted with. The solution is to assess the universality of candidate mechanisms in systematic cross-linguistic research, breaking free of the WEIRD ("Western, Educated, Industrialized, Rich, Democratic") bias (Henrich et al. 2010 BBS) that has plagued the science of human behavior and cognition. In my talk I will present recent work that explores the universal robustness of key mechanisms through the production data contained in corpora from very diverse languages. I will focus on four mechanisms: the production of (i) nouns vs verbs and (ii) variable affix orders and the effects that these mechanisms have on the evolution of morphology; the production of (iii) syntactic dependencies and (iv) information structures and the effects of these mechanisms on the evolution of word order patterns.

Short Bio

Balthasar Bickel got his graduate training in the Cognitive Anthropology Group at the Max-Planck Institute for Psycholinguistics in Nijmegen and received his Ph.D. degree in 1997 from the University of Zürich. From 1995 through 1998 he taught part-time in Zürich and worked as a researcher at the University of Mainz. He then spent three years at the University of California, Berkeley, on a postdoctoral fellowship sponsored by the Swiss National Science Foundation. In 2001 he completed his Habilitation at Zürich and was awarded an extracurricular professorship (Förderungsprofessur) by the Swiss National Science Foundation. From 2002 until 2011, he was professor of linguistic typology and variation at the University of Leipzig

(Department of Linguistics), and in 2011 he took over the chair of general linguistics at the University of Zürich. He is co-director (with Johanna Nichols) of the AUTOTYP research program for typological databasing. His core research question is what's where why in language. He wants to uncover the conditions that shape the spatio-temporal dynamics of language, and ultimately he seeks to understand their nature in the light of human evolution.

Sabine Stoll

Title: There are patterns in child-directed speech that make first language learning possible

One of the most amazing and unresolved puzzles in Cognitive Science is how children can learn any of the more than 7000 languages without ostensive instruction. Language learning is a long, implicit and step by step process relying on a number of domain general cognitive mechanism on the one hand and on an extremely powerful pattern detection mechanism on the other hand. Children learn their language from child-directed speech but languages differ extremely on all linguistic levels, from phonology, over morphology to semantics and syntax. So, the unresolved question is how children manage to learn all these diverse languages in the first 4 years of their life. Here I show that there are universally similar distributional patterns of linguistic units in child-directed speech which are optimally suited for learning. In Study 1 I show how these patterns enable children to learn an extremely complex verbal system of a polysynthetic language with over 4000 verb forms per verb. In Study 2 I show how 9 maximally diverse languages are structured to allow the extraction of parts of speech. In Study 3 on English, I show how child-directed speech is optimally suited for meaning extraction in comparing the distributional patterns of causative verbs in child-directed speech, adult-directed speech and written language.

Short Bio

Sabine Stoll is Professor for Psycholinguistics at the University of Zurich. After receiving her PhD at UC Berkeley she joined the Max Planck Institute for Evolutionary Anthropology in Leipzig where she worked for 10 years. Currently she is the head of the Psycholinguistics Lab at the Department of Comparative Linguistics and Co-Director of the Center for the Interdisciplinary Study of Language Evolution (ISLE) at the University of Zurich. Her main research focus is comparative first language development.

Isabelle Augenstein

Title: What can typological knowledge bases and language representations tell us about linguistic properties?

One of the core challenges in typology is to record properties of languages in a structured way. As a result of manual efforts, typological knowledge bases have emerged, which contains information about languages' phonological, morphological and syntactic properties; as well as information about language families. Ideally, such typological knowledge bases would provide useful information for multilingual NLP models to learn how to selectively share parameters.
A related area of research suggests a different way of encoding properties of languages, namely to learn language representation vectors directly from text documents.
In this talk, I will analyse and contrast these two ways of encoding linguistic properties, as well as present research on how the two can benefit one another.

Short Bio

Isabelle Augenstein is a tenure-track assistant professor at the University of Copenhagen, Department of Computer Science, where she heads the Copenhagen NLU research group. Her main research interests are weakly supervised and low-resource learning with applications

including information extraction, machine reading and fact checking. In the space of computational typology, she is interested in how information about properties of languages can aid multilingual learning, and has investigated language embeddings as well as automatically populating typological knowledge bases.