The mathematics of language learning

András Kornai
Computer and Automation Research Institute
Hungarian Academy of Sciences
andras@kornai.com

Gerald Penn
Department of Computer Science
University of Toronto
gpenn@cs.utoronto.edu

James Rogers
Computer Science Department
Earlham College
jrogers@cs.earlham.edu

Anssi Yli-Jyrä
Department of Modern Languages
University of Helsinki
anssi.yli-jyra@helsinki.fi

Over the past decade, attention has gradually shifted from the estimation of parameters to the learning of linguistic structure (for a survey see Smith 2011). The Mathematics of Language (MOL) SIG put together this tutorial, composed of three lectures, to highlight some alternative learning paradigms in speech, syntax, and semantics in the hopes of accelerating this trend.

Compounding the enormous variety of formal models one may consider is the bewildering range of ML techniques one may bring to bear. In addition to the surprisingly useful classical techniques inherited from multivariate statistics such as Principal Component Analysis (PCA, Pearson 1901) and Linear Discriminant Analysis (LDA, Fisher 1936), computational linguists have experimented with a broad range of neural net, nearest neighbor, maxent, genetic/evolutionary, decision tree, max margin, boost, simulated annealing, and graphical model learners. While many of these learners became standard in various domains of ML, within CL the basic HMM approach proved surprisingly resilient, and it is only very recently that deep learning techniques from neural computing are becoming competitive not just in speech, but also in OCR, paraphrase, sentiment analysis, parsing and vector-based semantic representations. The first lecture will provide a mathematical introduction to some of the fundamental techniques that lie beneath these linguistic applications of neural networks, such as: BFGS optimization, finite difference approximations of Hessians and Hessian-free optimization, contrastive divergence and variational inference.

Lecture 1: The mathematics of neural computing – *Penn*

Recent results in acoustic modeling, OCR, paraphrase, sentiment analysis, parsing and vector-based semantic representations have shown that natural language processing, like so many other corners of artificial intelligence, needs to pay more attention to neural computing.

I Gaussian Mixture Models
- Lagrange's theorem
- Stochastic gradient descent
- typical acoustic models using GMMs and HMMs

II Optimization theory
- Hessian matrices
- Broyden-Fletcher-Goldfarb-Shanno theory
- finite difference approximations of Hessians
- Hessian-free optimization
- Krylov methods

III Application: Product models
- products of Gaussians vs. GMMs
- products of "experts"
- Gibbs sampling and Markov-chain Monte Carlo
- contrastive divergence

IV Experimentation: Deep NNs for acoustic modeling
- intersecting product models with Boltzmann machines
- "generative pre-training"
- acoustic modeling with Deep Belief Networks
- why DBNs work well

V Variational inference
- variational Bayes for HMMs

In spite of the enormous progress brought by ML techniques, there remains a rather significant range of tasks where automated learners cannot yet get near human performance. One such is the unsupervised learning of word structure addressed by MorphoChallenge, another is the textual entailment task addressed by RTE.

The second lecture recasts these and similar problems in terms of learning weighted edges in a sparse graph, and presents learning techniques that seem to have some potential to better find spare finite state and near-FS models than EM. We will provide a mathematical introduction to the Minimum Description Length (MDL) paradigm and

spectral learning, and relate these to the better-known techniques based on (convex) optimization and (data-oriented) memorization.

Lecture 2: Lexical structure detection – *Kornai*

While modern syntactic theory focuses almost entirely on productive, rule-like regularities with compositional semantics, the vast bulk of the information conveyed by natural language, over 85%, is encoded by improductive, irregular, and non-compositional means, primarily by lexical meaning. Morphology and the lexicon provide a rich testing ground for comparing structure learning techniques, especially as inferences need to be based on very few examples, often just one.

I Motivation
- Why study structure?
- Why study lexical structure?

II Lexical structure
- Function words, content words
- Basic vocabulary (Ogden 1930, Swadesh 1950, Yasseri et al 2012)
- Estimation style

III Formal models of lexical semantics
- Associative (Findler 1979, Dumais 2005, CVS models)
- Combinatorial (FrameNet)
- Algebraic (Kornai 2010)

IV Spectral learning
- Case frames and valency
- Spectral learning as data cleaning (Ng 2001)
- Brew and Schulte im Walde 2002 (German), Nemeskey et al (Hungarian)
- Optionality in case frames

V Models with zeros
- Relating ungrammaticality and low probability (Pereira 2000, Stefanowitsch 2006)
- Estimation errors, language distances (Kornai 1998, 2011)
- Quantization error

VI Minimum description length
- Kolmogorov complexity and universal grammar (Clark 1994)
- MDL in morphology (Goldsmith 2000, Creutz and Lagus 2002, 2005,...)
- MDL for weighted languages
- Ambiguity
- Discarding data – yes, you can!
- Collapsing categories

VII New directions
- Spectral learning of HMMs (Hsu et al 2009, 2012)
- of weighted automata (Balle and Mohri 2012)
- Feature selection, LASSO (Pajkossy 2013)
- Long Short-Term Memory (Monner and Reggia 2012)
- Representation learning (Bengio et al 2013)

Given the broad range of competing formal models such as templates in speech, PCFGs and various MCS models in syntax, logic-based and association-based models in semantics, it is somewhat surprising that the bulk of the applied work is still performed by HMMs. A particularly significant case in point is provided by PCFGs, which have not proved competitive with straight trigram models. Undergirding the practical failure of PCFGs is a more subtle theoretical problem, that the nonterminals in better PCFGs cannot be identified with the kind of nonterminal labels that grammarians assume, and conversely, PCFGs embodying some form of grammatical knowledge tend not to outperform flatly initialized models that make no use of such knowledge. A natural response to this outcome is to retrench and use less powerful formal models, and the last lecture will be spent in the *subregular* space of formal models even less powerful than finite state automata.

Lecture 3: Subregular Languages and Their Linguistic Relevance – *Rogers* and *Yli-Jyrä*

The difficulty of learning a regular or context-free language in the limit from positive data gives a motivation for studying non-Chomskyan language classes. The lecture gives an overview of the taxonomy of the most important subregular classes of languages and motivate their linguistic relevance in phonology and syntax.

I Motivation
- Some classes of (sub)regular languages
- Learning (finite descriptions of) languages
- Identification in the limit from positive data
- Lattice leaners

II Finer subregular language classes
- The dot-depth hierarchy and the local and piecewise hierarchies
- k-Local and k-Piecewise Sets

III Relevance to phonology
- Stress patterns
- Classifying subregular constraints

IV Probabilistic models of language
- Strictly Piecewise Distributions (Heinz and Rogers 2010)

V Relevance to syntax
- Beyond the inadequate right-linear grammars
- Parsing via intersection and inverse morphism

- Subregular constraints on the structure annotations
- Notions of (parameterized) locality in syntax.

The relevance of some parameterized subregular language classes is shown through machine learning and typological arguments. Typological results on a large set of languages (Heinz 2007, Heinz et al 2011) relate language types to the theory of subregular language classes.

There are finite-state approaches to syntax showing subregular properties. Although structure-assigning syntax differs from phonotactical constraints, the inadequacy of right-linear grammars does not generalize to all finite-state representations of syntax. The linguistic relevance and descriptive adequacy are discussed, in particular, in the context of intersection parsing and conjunctive representations of syntax.

Instructors

Anssi Yli-Jyrä is Academy Research Fellow of the Academy of Finland and Visiting Fellow at Clare Hall, Cambridge. His research focuses on finite-state technology in phonology, morphology and syntax. He is interested in weighted logic, dependency complexity and machine learning.

James Rogers is Professor of Computer Science at Earlham College, currently on sabbatical at the Department of Linguistics and Cognitive Science, University of Delaware. His primary research interests are in formal models of language and formal language theory, particularly model-theoretic approaches to these, and in cognitive science.

Gerald Penn teaches computer science at the University of Toronto, and is a Senior Member of the IEEE. His research interests are in spoken language processing, human-computer interaction, and mathematical linguistics.

András Kornai teaches at the Algebra Department of the Budapest Institute of Technology, and leads the HLT group at the Computer and Automation Research Institute of the Hungarian Academy of Sciences. He is interested in everything in the intersection of mathematics and linguistics. For a list of his publications see http://kornai.com/pub.html.

Online resources

Slides for the tutorial:
http://molweb.org/acl13tutorial.pdf
Bibliography:
http://molweb.org/acl13refs.pdf
Software:
http://molweb.org/acl13sw.pdf