Language Technology: Research and Development

Language Technology Research and Development

Joakim Nivre

Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se
The Name of the Game

Computational Linguistics (CL)

Natural Language Processing (NLP)

[Human] Language Technology ([H]LT)

[Natural] Language Engineering ([N]LE)
The Name of the Game

Computational Linguistics (CL)
- Study of natural language from a computational perspective

Natural Language Processing (NLP)
- Study of computational models for processing natural language

[Human] Language Technology ([H]LT)
- Development and evaluation of applications based on CL/NLP

[Natural] Language Engineering ([N]LE)
- Same as [H]LT but obsolete?
An Interdisciplinary Field

Linguistics
  ▶ Theory, language description, data analysis (annotation)

Computer science
  ▶ Theory, data models, algorithms, software technology

Mathematics
  ▶ Theory, abstract models, analytic and numerical methods

Statistics
  ▶ Theory, statistical learning and inference, data analysis
Linguistics

- Structuralist linguistics (1915–1960)
  - Language as a network of relations (phonology, morphology)
  - Inductive discovery procedures
- Generative grammar (1960–)
  - Language as a generative system (syntax)
  - Deductive formal systems (formal language theory)
  - NLP systems based on linguistic theories
Linguistics

- Recent trends (1990–):
  - Language processing (psycholinguistics, neurolinguistics)
  - Strong empiricist movement (corpus linguistics)
  - NLP systems based on linguistically annotated data

- Theoretical and computational linguistics have diverged

Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous? (Workshop at EACL 2009)
Computer Science

- Theoretical computer science
  - Turing machines and computability (Church-Turing thesis)
  - Algorithm and complexity theory (cf. formal language theory)
- Artificial Intelligence
  - Early work on symbolic logic-based systems (GOFAI)
  - Trend towards machine learning and sub-symbolic systems
  - Parallel development in natural language processing
Mathematics

- Mathematical model
  - Description of real-world system using mathematical concepts
  - Formed by abstraction over real-world system
  - Provide computable solutions to problems
  - Solutions interpreted and evaluated in the real world
- Mathematical modeling fundamental to (many) science(s)
Mathematics

- Real-world language technology problem:
  - Syntactic parsing: sentence $\Rightarrow$ syntactic structure
  - No precise definition of relation from inputs to outputs
  - At best annotated data samples (treebanks)

- Mathematical model:
  - Probabilistic context-free grammar $G$
    \[
    T^* = \arg\max_{T: \text{yield}(S) = T} P_G(T)
    \]
    - $T^*$ can be computed exactly in the model
    - $T^*$ may or may not give a solution to the real problem

- How do we determine whether a model is good or bad?
Statistics

Probability theory
- Mathematical theory of uncertainty

Descriptive statistics
- Methods for summarizing information in large data sets

Statistical inference
- Methods for generalizing from samples to populations
Statistics

- Probability theory
  - Framework for mathematical modeling
  - Standard models: HMM, PCFG, Naive Bayes
- Descriptive statistics
  - Summary statistics in exploratory empirical studies
  - Evaluation metrics in experiments (accuracy, precision, recall)
- Statistical inference
  - Estimation of model parameters (machine learning)
  - Hypothesis testing about systems (evaluation)
Language Technology R&D

Sections in *Transactions of the ACL (TACL)*:

- Theoretical research
- Empirical research
- Applications and tools
- Resources and evaluation
Language Technology R&D

Sections in Transactions of the ACL (TACL):

- Theoretical research – deductive approach
- Empirical research – inductive approach
- Applications and tools – design and construction
- Resources and evaluation – data and method
Theoretical Research

- Formal theories of language and computation
- Studies of models and algorithms in themselves
- Claims justified by formal argument (deductive proofs)
- Often implicit relation to real-world problems and data
Theoretical Research

Efficient Parsing for Head-Split Dependency Trees.
Transactions of the Association for Computational Linguistics 1, 267–278.

- Contribution:
  - Parsing algorithms for non-projective dependency trees
  - Added constraints reduce complexity from $O(n^7)$ to $O(n^5)$

- Approach:
  - Formal description of algorithms
  - Proofs of correctness and complexity
  - No implementation or experiments
  - Empirical analysis of coverage after adding constraints
Empirical Research

- Empirical studies of language and computation
- Studies of models and algorithms applied to data
- Claims justified by experiments and statistical inference
- Explicit relation to real-world problems and data
Empirical Research


▶ Contribution:
  ▶ Latent variable CRFs for unsupervised part-of-speech tagging
  ▶ Learning from both type and token constraints

▶ Approach:
  ▶ Formal description of mathematical model
  ▶ Statistical inference for learning and evaluation
  ▶ Multilingual data sets used in experiments
Applications and Tools

- Design and construction of LT systems
- Primarily end-to-end applications (user-oriented)
- Claims often justified by proven experience
- May include experimental evaluation or user study
Applications and Tools


- **Contribution:**
  - In-depth description of design and application development
  - Extensive evaluation in the context of application (real users)

- **Approach:**
  - Case study – concrete instance in context
  - Semi-formal system description (flowcharts, examples)
  - Statistical inference for evaluation
Resources and Evaluation

Resources
- Collection and annotation of data (for learning and evaluation)
- Design and construction of knowledge bases (grammars, lexica)

Evaluation
- Protocols for (empirical) evaluation
  - Intrinsic evaluation – task performance
  - Extrinsic evaluation – effect on end-to-end application
- Methodological considerations:
  - Selection of test data (sampling)
  - Evaluation metrics (intrinsic, extrinsic)
  - Significance testing (statistical inference)
Creating a Live, Public Short Message Service Corpus:

- Contribution:
  - Free SMS corpus in English and Chinese (> 70,000 msgs)
  - Discussion of methodological considerations

- Approach:
  - Crowdsourcing using mobile phone apps
  - Automatic anonymization using regular expressions
  - Linguistic annotation as future plans
Language Technology as a Science

- Scientific reasoning
  - Deduction common in theoretical research
  - Induction underlies machine learning and statistical evaluation
  - Inference to the best explanation in experimental studies

- Scientific explanation
  - Explanations based on general laws are rare
  - Explanations based on statistical generalizations are the norm

- Reproducibility/replicability
  - Important in theory but problematic in practice
  - Recent initiatives to publish data and software with papers

Language Technology as a Science

- The “empirical revolution” in language technology
  - Before 1990: Rationalist approaches and qualitative evaluation
  - Today: Empirical approaches and quantitative evaluation
- What happened?
  - Paradigm shift in Kuhn’s sense?
  - Just another swing of the pendulum?
  - Language technology becoming a mature science?
Ethics in Language Technology

- Increasing attention in the (larger) community
- Some issues raised by Hovy and Spruit:
  - Exclusion – data bias
  - Overgeneralization – modeling bias
  - Dual-use problems
- First Workshop on Ethics in NLP held in 2017