Introduction to NLP

Background
Linguistic Knowledge

• Constituents:
  – Children eat pizza.
  – They eat pizza.
  – My cousin’s neighbor’s children eat pizza.
  – Eat pizza!

• Collocations:
  – Strong beer but *powerful beer
  – Big sister but *large sister
  – Stocks rise but ?stocks ascend
    • in the past: 225,000 hits vs. 47 hits on Google, now 550,000 vs 57,000

• How to get this knowledge in the system:
  – Manual rules
  – Automatically acquired from large text collections (corpora)
Linguistic Knowledge

• Knowledge about language:
  – Phonetics and phonology – the study of sounds
  – Morphology – the study of word components
  – Syntax – the study of sentence and phrase structure
  – Lexical semantics – the study of the meanings of words
  – Compositional semantics – how to combine words
  – Pragmatics – how to accomplish goals
  – Discourse conventions – how to deal with units larger than utterances

• Separate lecture
Finite-state Automata
Theoretical Computer Science

• Automata
  – Deterministic and non-deterministic finite-state automata
  – Push-down automata

• Grammars
  – Regular grammars
  – Context-free grammars
  – Context-sensitive grammars

• Complexity

• Algorithms
  – Dynamic programming
Mathematics and Statistics

- Probabilities
- Statistical models
- Hypothesis testing
- Linear algebra
- Optimization
- Numerical methods
Mathematical and Computational Tools

- Language models
- Estimation methods
- Context-free grammars (CFG)
  - for trees
- Hidden Markov Models (HMM)
  - for sequences
- Conditional Random Fields (CRF)
- Generative/discriminative models
- Maximum entropy models
Statistical Techniques

- Vector space representation for WSD
- Noisy channel models for MT
- Graph–based Random walk methods for sentiment analysis

\[
\hat{E} = \arg\max_{E \in \text{English}} P(E \mid F) \\
= \arg\max_{E \in \text{English}} \frac{P(F \mid E)P(E)}{P(F)} \\
= \arg\max_{E \in \text{English}} P(F \mid E)P(E)
\]
Artificial Intelligence

• Logic
  – First-order logic
  – Predicate calculus

• Agents
  – Speech acts

• Planning

• Constraint satisfaction

• Machine learning
NLP