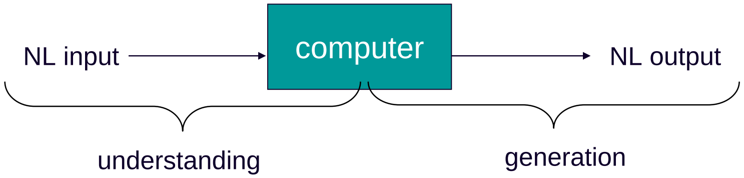
Last class: Why study NLP?



NL input

computer

NL output

understanding

generation

– Useful applications
– Interdisciplinary
– Challenging

Topics for Today

- Why is NLP a challenging area of research?
- Brief history of NLP
- Writing critiques

Why is NLP such a difficult problem?

Ambiguity!!!! …at **all** levels of analysis ☹

- Phonetics and phonology
  - Concerns how words are related to the sounds that realize them
  - Important for speech-based systems.
    - » "I scream" vs. "ice cream"
    - » "nominal egg"
  - Moral is:
    - » It's very hard to recognize speech.
    - » It's very hard to wreck a nice beach.
- Morphology
  - Concerns how words are constructed from sub-word units
  - Unionized
    - » un-ionized in chemistry?

Why is NLP such a difficult problem?

Ambiguity!!!! …at **all** levels of analysis ☹

- Syntax
  - Concerns sentence structure
  - Different syntactic structure implies different interpretation
    - » Squad helps dog bite victim.
      - ◆[$_{\text{np}}$ squad] [$_{\text{vp}}$ helps [$_{\text{np}}$ dog bite victim]]
      - ◆[$_{\text{np}}$ squad] [$_{\text{vp}}$ helps [$_{\text{np}}$ dog] [$_{\text{inf-clause}}$ bite victim]]
    - » Helicopter powered by human flies.
    - » Visiting relatives can be trying.

Why is NLP such a difficult problem?

Ambiguity!!!! …at **all** levels of analysis ☹

- Semantics
  - Concerns what words mean and how these meanings combine to form sentence meanings.
    - Jack invited Mary to the Halloween **ball**.
      - dance vs. some big sphere with with Halloween decorations?
    - Visiting relatives can be trying.
    - Visiting museums can be trying.
      - Same set of possible syntactic structures for this sentence
      - But the meaning of **museums** makes only one of them plausible

Why is NLP such a difficult problem?

Ambiguity!!!! …at **all** levels of analysis ☹

- Discourse
  - Concerns how the immediately preceding sentences affect the interpretation of the next sentence
    - Merck & Co. formed a joint venture with Ache Group, of Brazil. **It** will be called Prodome Ltd.
    - Merck & Co. formed a joint venture with Ache Group, of Brazil. **It** will own 50% of the new company to be called Prodome Ltd.
    - Merck & Co. formed a joint venture with Ache Group, of Brazil. **It** had previously teamed up with Merck in two unsuccessful pharmaceutical ventures.

Why is NLP such a difficult problem?

Ambiguity!!!! …at **all** levels of analysis ☹

- Pragmatics
  - Concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

    ``I just came from New York.''

    - Would you like to go to New York today?
    - Would you like to go to Boston today?
    - Why do you seem so out of it?
    - Boy, you look tired.

Early Roots: 1940's and 1950's

- Work on two foundational paradigms
  - Automaton
    - Turing's (1936) model of algorithmic computation
    - Kleene's (1951, 1956) finite automata and regular expressions
    - Shannon (1948) applied probabilistic models of discrete Markov processes to automata for language
    - Chomsky (1956)
      - First considered finite-state machines as a way to characterize a grammar
      - Led to the field of formal language theory

Early Roots: 1940's and 1950's

- Work on two foundational paradigms
  - – Probabilistic or information-theoretic models for speech and language processing
    - Shannon: the "noisy channel" model
    - Shannon: borrowing of "entropy" from thermodynamics to measure the information content of a language

Two Camps: 1957-1970

- Symbolic paradigm
  - – Chomsky
    - » Formal language theory, generative syntax, parsing
    - » Linguists and computer scientists
    - » Earliest complete parsing systems
      - ◆Zelig Harris, UPenn
      - ◆…A possible critique reading!!

Two Camps: 1957-1970

- Symbolic paradigm
  - – Artificial intelligence
    - » Created in the summer of 1956
    - » Two-month workshop at Dartmouth
    - » Focus of the field initially was the work on reasoning and logic (Newell and Simon)
    - » Early natural language systems were built
      - ◆Worked in a single domain
      - ◆Used pattern matching and keyword search

Two Camps: 1957-1970

- Stochastic paradigm
    - » Took hold in statistics and EE
    - » Late 50's: applied Bayesian methods to OCR
    - » Mosteller and Wallace (1964): applied Bayesian methods to the problem of authorship attribution for *The Federalist* papers.

Additional Developments

- 1960's
  - First serious testable psychological models of human language processing
    - Based on transformational grammar
  - First on-line corpora
    - The Brown corpus of American English
      - 1 million word collection
      - Samples from 500 written texts
      - Different genres (news, novels, non-fiction, academic,….)
      - Assembled at Brown University (1963-64, Kucera and Francis)
    - William Wang's (1967) DOC (Dictionary on Computer)
      - On-line Chinese dialect dictionary

1970-1983

- Explosion of research
  - Stochastic paradigm
    - Developed speech recognition algorithms
      - HMM's
      - Developed independently by Jelinek et al. at IBM and Baker at CMU
  - Logic-based paradigm
    - Prolog, definite-clause grammars (Pereira and Warren, 1980)
    - Functional grammar (Kay, 1979) and LFG

1970-1983

- Explosion of research
  - Natural language understanding
    - SHRDLU (Winograd, 1972)
    - The Yale School
      - Focused on human conceptual knowledge and memory organization
    - Logic-based LUNAR question-answering system (Woods, 1973)
  - Discourse modeling paradigm

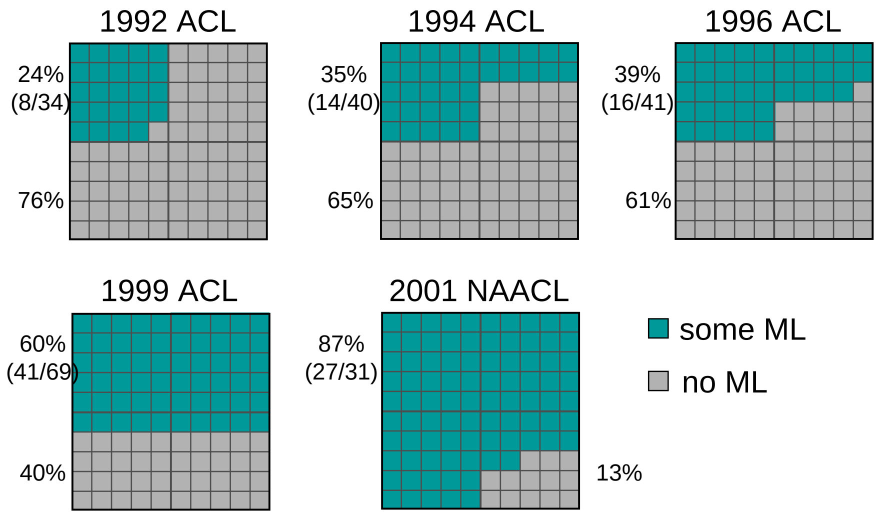
Revival of Empiricism and FSM's

- 1983-1993
  - Finite-state models
    - Phonology and morphology (Kaplan and Kay, 1981)
    - Syntax (Church, 1980)
  - Return of empiricism
    - Rise of probabilistic models in speech and language processing
    - Largely influenced by work in speech recognition at IBM
  - Considerable work on natural language generation

A Reunion of a Sort…

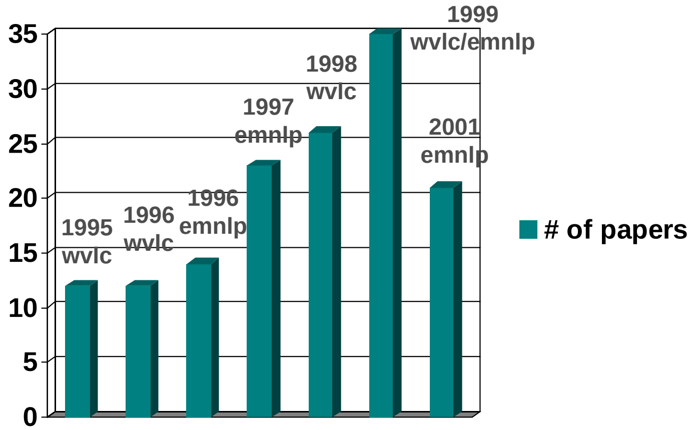
- 1994-1999
  - Probabilistic and data-driven models had become quite standard
  - Increases in speed and memory of computers allowed commercial exploitation of speech and language processing
    - Spelling and grammar checking
  - Rise of the Web emphasized the need for language-based information retrieval and information extraction

Statistical and Machine Learning Approaches Rule!

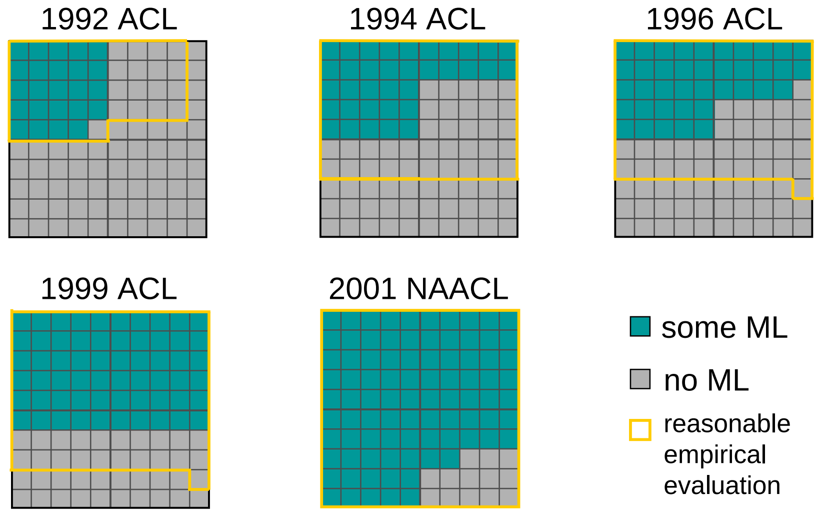


WVLC and EMNLP Conferences

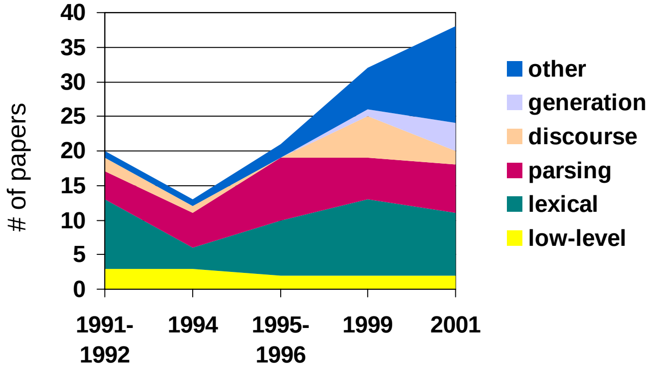
- Workshop on Very Large Corpora
- Conference on Empirical Methods in NLP



Empirical Evaluation



Progression of NL learning tasks



Critique Guidelines

- <=1 page, typed (single space)
- The purpose of a critique is **not** to summarize the paper; rather you should choose one or two points about the work that you found interesting.
- Examples of questions that you might address are:
  - What are the strengths and limitations of its approach?
  - Is the evaluation fair? Does it achieve it support the stated goals of the paper?
  - Does the method described seem mature enough to use in real applications? Why or why not? What applications seem particularly amenable to this approach?
  - What good ideas does the problem formulation, the solution, the approach or the research method contain that could be applied elsewhere?
  - What would be good follow-on projects and why?

Critique Guidelines

  - Are the paper's underlying assumptions valid?
  - Did the paper provide a clear enough and detailed enough description of the proposed methods for you to be able to implement them? If not, where is additional clarification or detail needed?

- Avoid **unsupported** value judgments, like ``I liked...'' or ``I disagreed with...'' If you make judgments of this sort, explain why you liked or disagreed with the point you describe.
- Be sure to distinguish comments about the writing of the paper from comment about the technical content of the work.