Language and Statistics II

Lecture 1: Introduction

Noah Smith

Today's Plan

- What's this course about?
  - Goals
  - Topics
- How will we be evaluated?
- Some history
- Q&A

My Goals

- You'll **read**, **write**, and **present** technical information better.
- A deeper, more **connected** understanding of the EMNLP literature.
- New ideas all around for bringing NLP "tasks" closer to NLP **applications**.
- You'll refine your taste for good research.
- You'll enjoy the next ACL/NAACL/EMNLP and impress people with your insight and communication skills!

“Who Am I and Why Am I Here?”

(introductions all around)

Presumptions

- You took L&S I, or equivalent
- You believe in statistical methods for language technology
- You want to know more about current practice in the area

Topics

- In lectures, we'll mostly cover tools and "tasks"
- Your lit review will cover an application
- If you've taken ML, this course may feel applications-focused.
- If you've taken applied NLP courses (ASR, IR, MT, IE, etc.), this course may feel theoretical.
- If you've taken both, this course will tie things together.
- If you've taken linguistics, this course will be frustrating - that's a good thing!

Topics

- Sequence models (might be review)
  - Markov models
  - HMMs
  - Algorithms and applications

Topics

- Sequence models (might be review)
- Log-linear models
  - Theory
  - Practice
  - Examples
    - Ratnaparkhi tagger
    - CRFs
  - Tasks

Topics

- Sequence models (might be review)
- Log-linear models
- Weighted finite-state technology
  - Algorithms
  - Applications
  - Tools

Topics

- Sequence models (might be review)
- Log-linear models
- Weighted finite-state technology
- Weighted grammars and parsing
  - Theory
  - Eisner, Charniak, Ratnaparkhi, Collins, McDonald
  - Maybe: LTAG, CCG
  - Practical issues

Topics

- Sequence models (might be review)
- Log-linear models
- Weighted finite-state technology
- Weighted grammars and parsing
- Dynamic programming
  - Unified framework for aligning, labeling, parsing, …
  - Implementation challenges
  - Limitations

Topics

- Sequence models (might be review)
- Log-linear models
- Weighted finite-state technology
- Weighted grammars and parsing
- Dynamic programming
- Going discriminative
  - Blast from the past: transformation-based learning
  - Perceptrons and maximum-margin training
  - Reranking

Topics

- Sequence models (might be review)
- Log-linear models
- Weighted finite-state technology
- Weighted grammars and parsing
- Dynamic programming
- Going discriminative
- Going unsupervised
  - Expectation-Maximization
  - Contrastive Estimation
  - Dirichlet processes (maybe)

Topics

- Sequence models (might be review)
- Log-linear models
- Weighted finite-state technology
- Weighted grammars and parsing
- Dynamic programming
- Going discriminative
- Going unsupervised
- Going semi-supervised
  - Self-training
  - Yarowsky algorithm, Cotraining

Topics

- Sequence models (might be review)
- Log-linear models
- Weighted finite-state technology
- Weighted grammars and parsing
- Dynamic programming
- Going discriminative
- Going unsupervised
- Going semi-supervised
- Time/interest depending: MT, OT, Kernels

Evaluation

- Lectures, suggested readings ➔
  - ~4-6 Assignments (20%)
  - Final Exam (20%)

- Literature review
  - Written document (35%)
  - Oral presentation (25%)

Literature Review

Comprehensive review of the literature:

- Define **clearly** a problem within NLP
- Define existing evaluation procedures
- Discuss available datasets
- Thorough, coherent discussion of existing techniques
- Comparison among techniques, if possible
- Current **obstacles**
- **Insights** on tackling or avoiding those obstacles, improving evaluation, “scaling up,” etc.

Suggested Topics

- Question answering
- Textual entailment and paraphrase
- Morphology induction and modeling
- Syntax-based machine translation
- Data-oriented parsing and translation
- Syntax-based language modeling
- Finite-state parsing
- Optimality theory

(You're welcome to propose other areas!)

Carrots

- Theses usually have literature reviews.
- *Computational Linguistics* will start publishing literature reviews soon.
- Well-written reviews, when put online, tend to be oft-referenced and oft-cited.

Deliverables

- Sept. 12: pick topics, initial reading list
- Oct. 16-20: progress meeting
- Nov. 10: first draft
- Last 1-2 weeks of class: talks
- Dec. 8: final version

Question

- **Interspeech** is the fourth week of term; who intends to be there?

Supplications

- New faculty member, new course …
  - Please have patience!
  - All feedback is welcome!
- Ask questions!
  - I don’t know everything.
  - But I probably know where to look or who to ask.

(Most of)
The Rest (of the Lecture)
is History

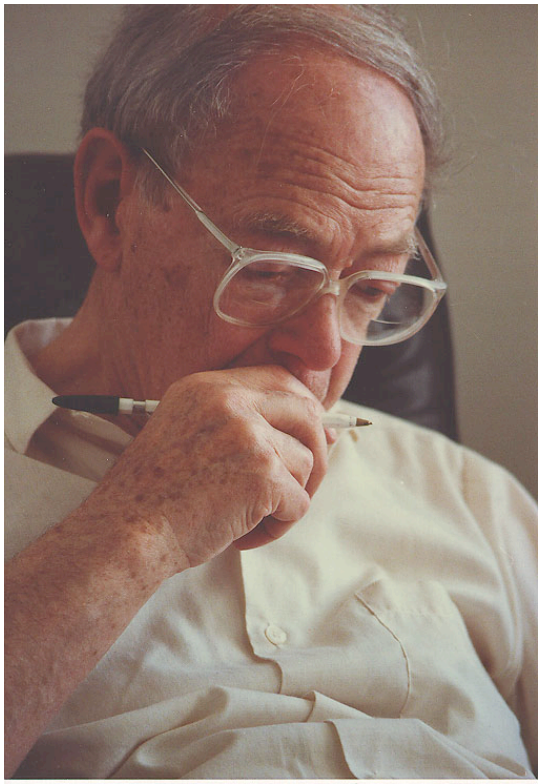
Let’s not repeat the mistakes of the past.

Cocktail party conversation at the next ACL.

Issues to keep in mind as we proceed.

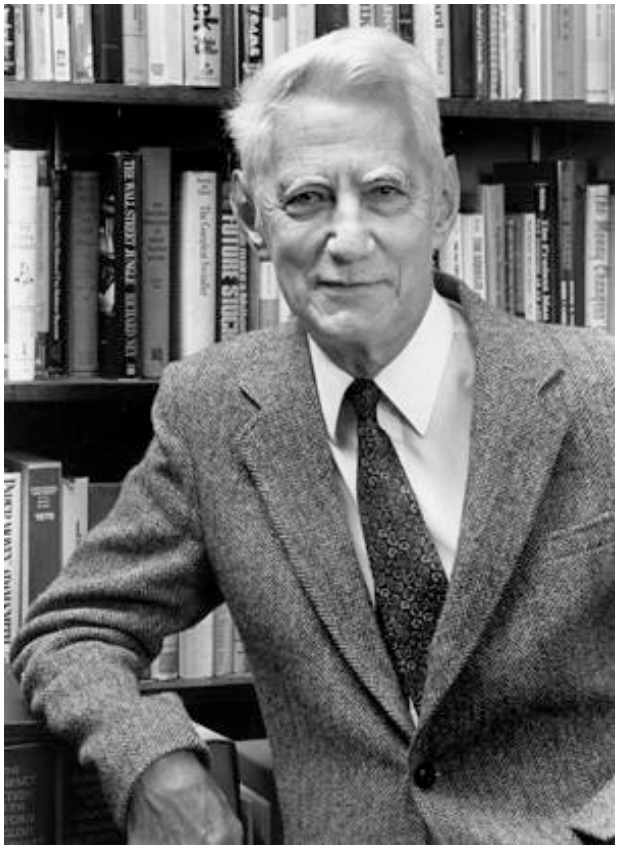
Zellig Harris (1909-1992)

- Validation criteria for linguistic analysis
- Linguistic transformations as a tool for describing language mathematically
- Centrality of data!
- Students: Chomsky, Gleitman, Joshi, …
- Note: structuralism never died in Europe.



Claude Shannon (1916-2001)

- Father of information theory
- Entropy: a mathematical measure of uncertainty
- Information can be encoded digitally; questions include how to encode information **efficiently** and **reliably**.
- Huge impact on speech recognition (and space exploration and digital media invention and …)
- 1949: Weaver compared translation to cryptography

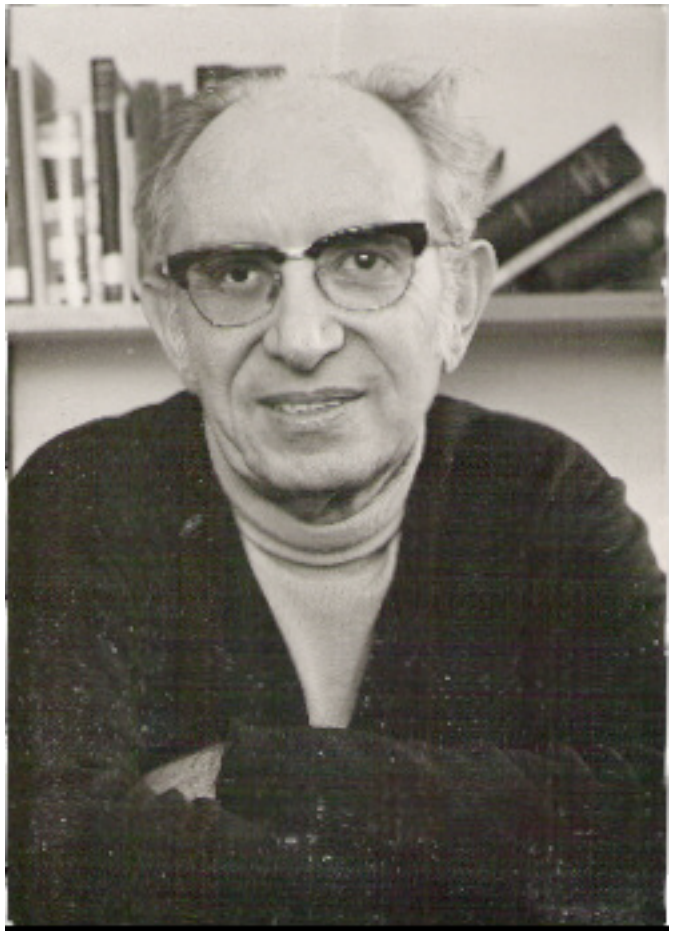


Victor Yngve (1920-)

- Early computational linguist
- Showed “depth limit” of human sentence processing - restricted left branching (but not right)
- Theme: what are the real observables in language study? Sound waves!
- Early programming language, COMIT, for linguists (influenced SNOBOL)

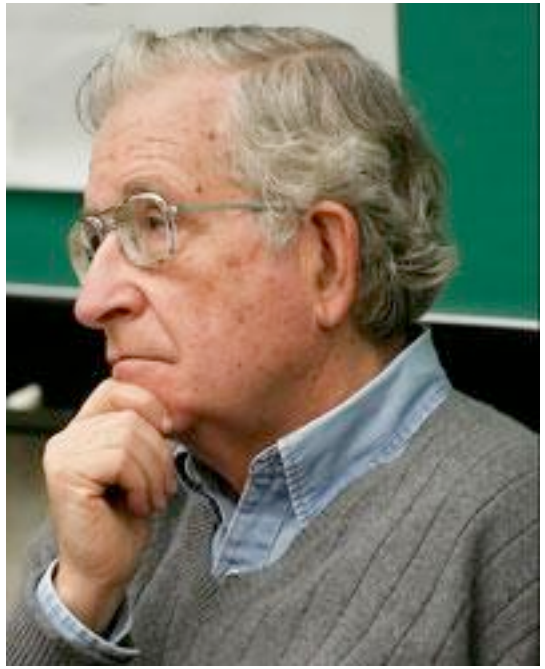
Yehoshua Bar-Hillel (1915-1975)

- First academic to work on MT
- Believed in the close relationship of logic and language
- Tremendous foresight in identifying the problems in MT … before it existed.



Noam Chomsky (1928-)

- Universal grammar ➔ productivity
- Chomsky hierarchy
- Generative grammar (P&P, GB, Minimalism), a series of (mainly syntax) theories that are based largely on the grammatical/ungrammatical boundary.
- Data? Native speaker judgments.
- Claim: "probability of a sentence" is a meaningless idea.



ALPAC Report

- “Automatic Language Processing Advisory Committee” (1964-6)
- Skeptical of MT research; to paraphrase, “we don’t have it and it looks like we never will.”
- Supportive of basic research in linguistics: “we need understanding!”
- Bar-Hillel left the field
- Killed MT for a while

The Rise of Rationalism (1960-85)

- In linguistics, more and more focus on **syntax**, less on processing & algorithms
- Rule-based approaches in AI ≈ innate knowledge of language (reasoning, etc.)
- Many linguists didn't/don't care about applications

Science-Engineering Debate

- NLP = CL?
- Are we doing science or engineering?
- Can computational experiments tell us anything about human intelligence?
- Can theories of human intelligence give insight to engineering problems?
- Beware the worst of both worlds:
  - Science requires no application …
  - Engineering requires no rigor …

The Return of Empiricism (1985-today)

- Late 1980s: ASR meets NLP
- Major efforts at IBM
  - Also, AT&T, U. Penn, CMU
- Candide - statistical MT model
- Spatter - statistical parser
- Now empirical methods are mainstream

Rationalist-Empiricist Debate

- Visible across AI, Cognitive Science, Linguistics
- Skinner/Chomsky
- ASR/GOFAI
- Connectionism/Symbolic systems
- Corpus-based linguistics/“theoretical” linguistics
- Statistical NLP/Knowledge-based NLP

- Mature view: science (many unresolved questions) vs. engineering (use what you’ve got)

Some Opinions

- Good engineering is
  - Intuitive and understandable
  - Formally rigorous
  - Replicable (like good science!)
- Mediocre engineering can often be "cleaned up" later (if replicable)
- Linguistics (& cognitive science) ➔ smarter models and features
- Empirical NLP has **more** to offer to linguistics than rule-based NLP!
- NLP is one of the most interesting & difficult ML application areas