Structured Prediction for Language and Other Discrete Data (10-710 and 11-763)

Introductory Lecture

A Little Bit of History

1935: Zipf's law

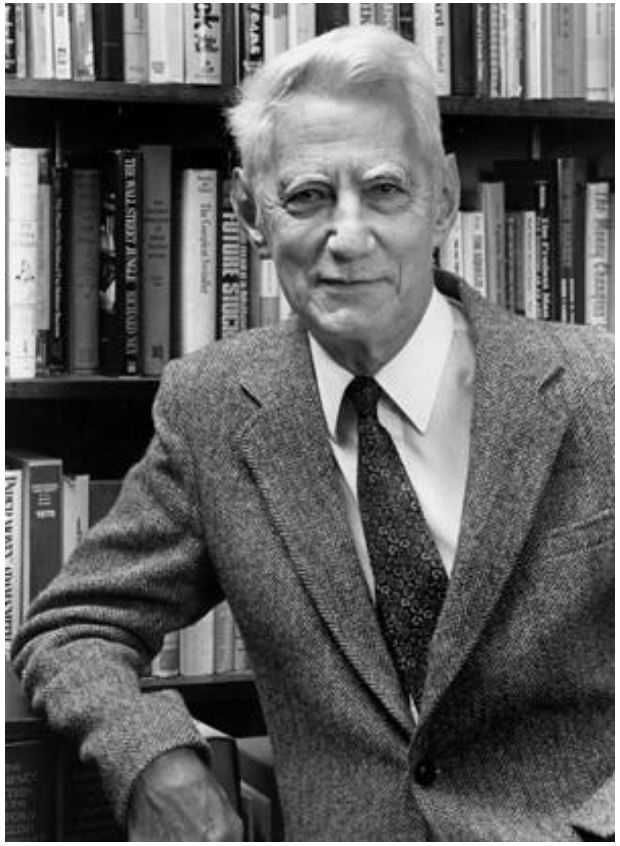
1940s & 1950s: empiricism: Shannon, Weaver, Harris, Yngve

George Kingsley Zipf, 1935

$$p(w) \approx \frac{1}{rank(w)}$$

- Heavy tail in word distributions
- (Incomes, too; accurately predicted revolution in Indonesia)

Claude Shannon, 1948

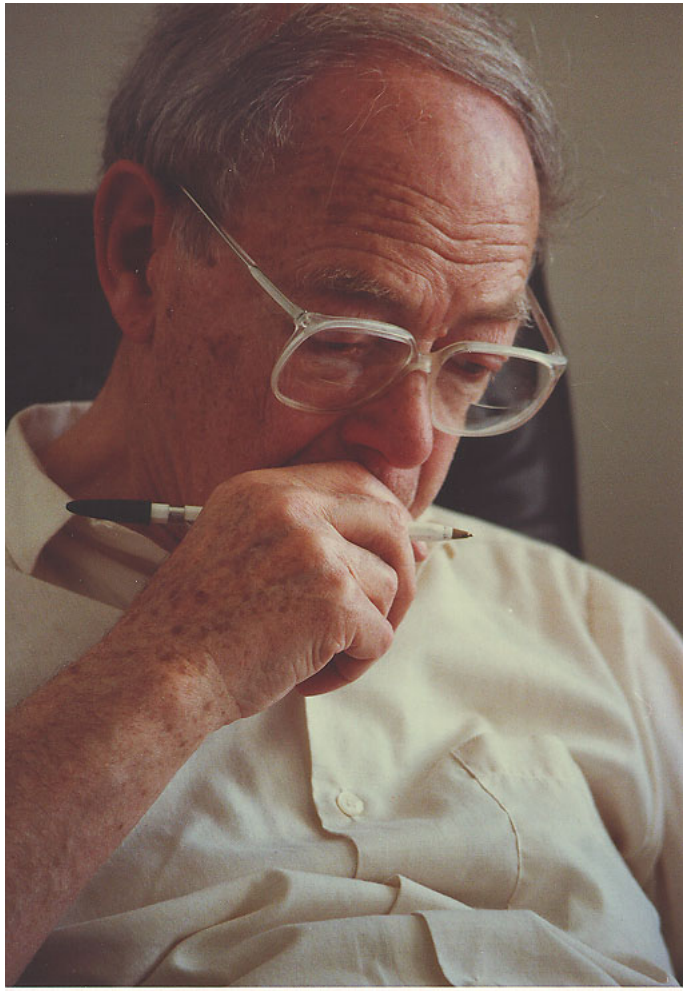


- Father of information theory
- Entropy: a mathematical measure of uncertainty
- Information can be encoded digitally; questions include how to encode information efficiently and reliably.
- Huge impact on speech recognition (and space exploration and digital media invention and …)

Warren Weaver, 1949

- “One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”

Zellig Harris, 1940s and forward



- Centrality of data for linguistic analysis
- Transformations (a step toward computational models of language)
- Heavy use of mathematics in linguistics

Victor Yngve, 1958

- 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)
- Random sentence generation (in the 1950s)

A Little Bit of History

1935: Zipf's law

1940s & 1950s: empiricism: Shannon, Weaver, Harris, Yngve

1960-1985: rationalism/representations/formalisms/syntax/unapplied AI

- 1962: ACL (then MTACL) begins
- 1964-6: ALPAC report, MT winter, Bar-Hillel leaves the field

1980: ICML begins

~1985: statistical and information theoretic methods catch hold again in NLP, in part due to their success in ASR

- This has continued unabated for 25+ years, with help from Moore's Law-type phenomena

1986: LTI founded (then called "CMT")

1993: "Very Large Corpora" workshops start at ACL

1996: EMNLP conference starts

~1997: Lafferty and Rosenfeld start teaching "Language and Statistics" at CMU

1998-early 2000s: Internet boom, commercial language technologies becoming viable

~2003: MLD founded (then called "CALD")

2004: Cohen starts teaching "Information Extraction"

2006: Smith starts teaching "Language and Statistics 2"

2011: Cohen and Smith start teaching "Structured Prediction"

What is Structured Prediction?

Having observed some information (input) …

- Binary classification: predict a coin toss (given some information)
- Multi-class: predict which side of a die (given some information)
- Structured prediction: choose among a very large number of complex outcomes.
  - Large means "exponential in the size of the input."

E.g., (Part of Speech) Tagging

E.g., Segmentation *into* Words

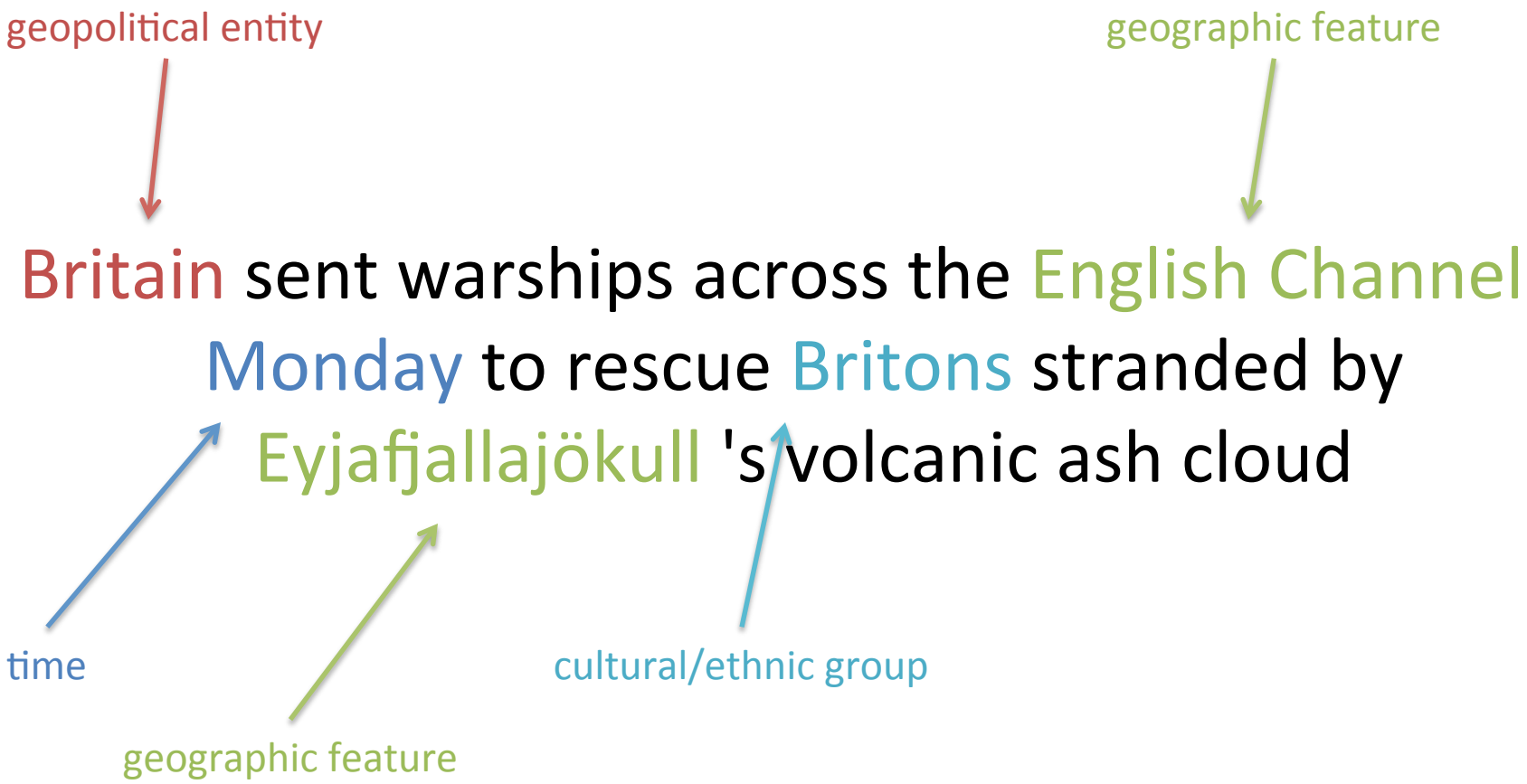
第二阶段的奥运会体育比赛门票与残奥会开闭幕式门票的预订工作已经结束,现在进入门票分配阶段。在此期间,我们不再接受新的门票预订申请。

E.g., Segmentation *within* Words

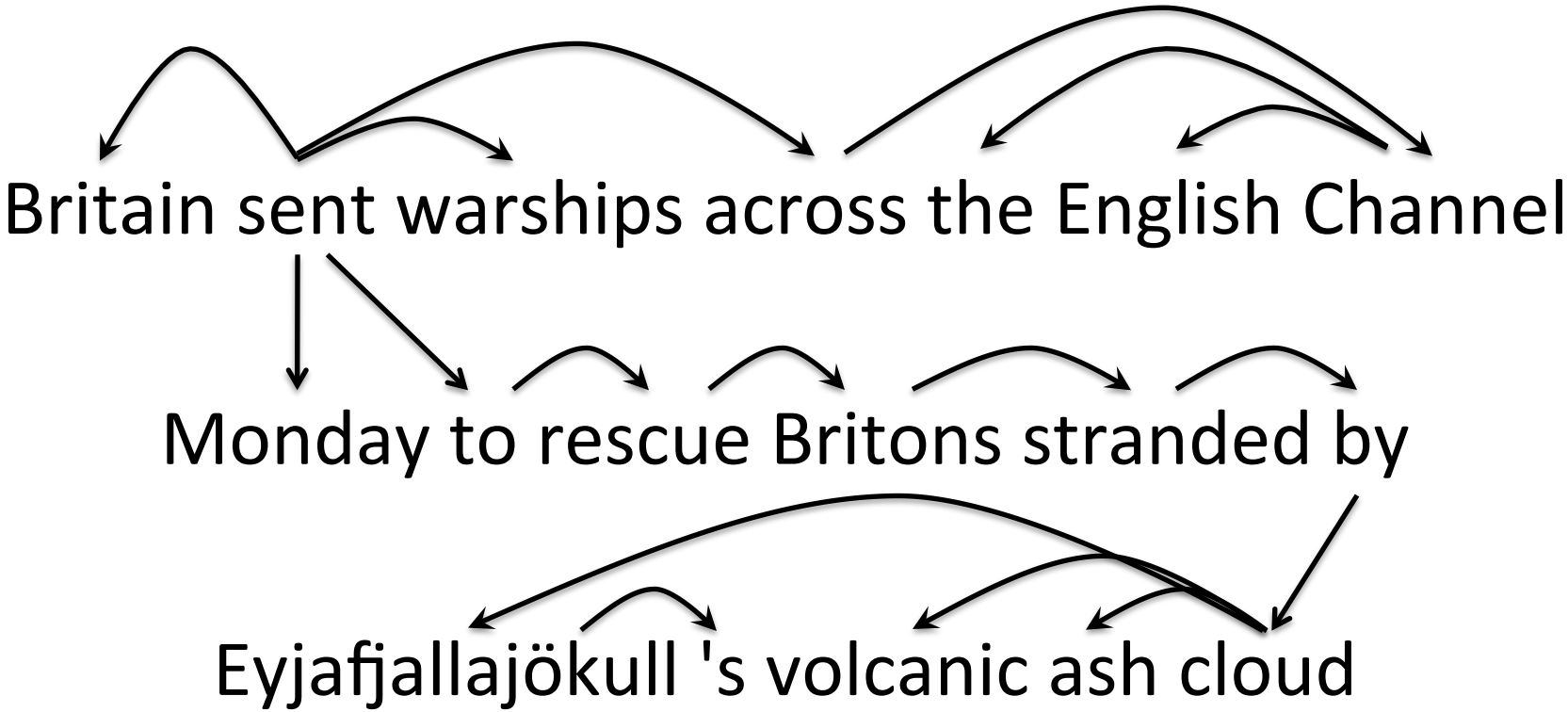
uygarlaştıramadıklarımızdanmışsınızcasına

“(behaving) as if you are among those whom we could not civilize”

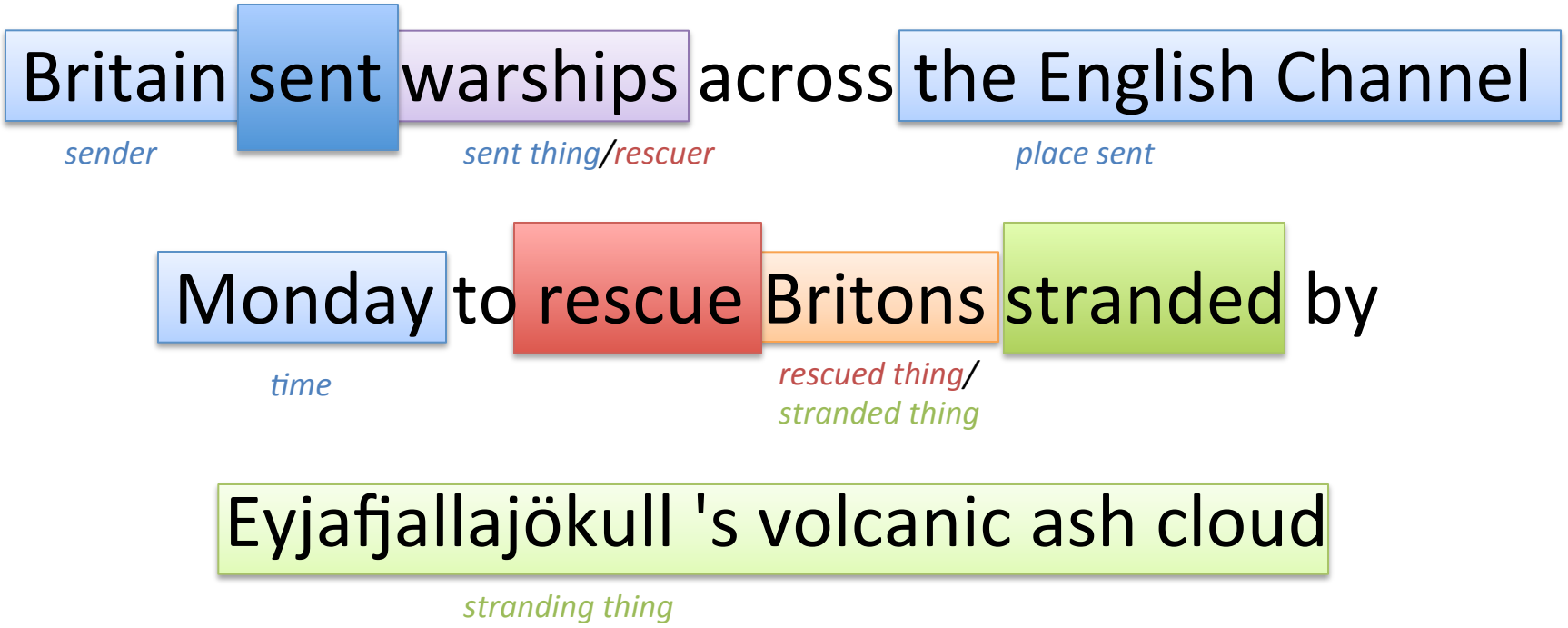
E.g., Segmentation *and* Tagging



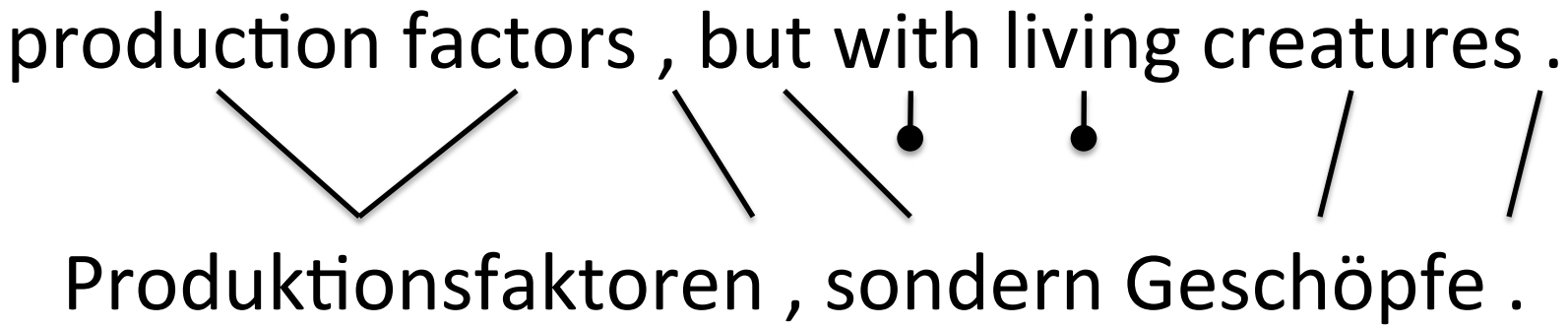
E.g., Trees


Britain sent warships across the English Channel
Monday to rescue Britons stranded by
Eyjafjallajökull 's volcanic ash cloud

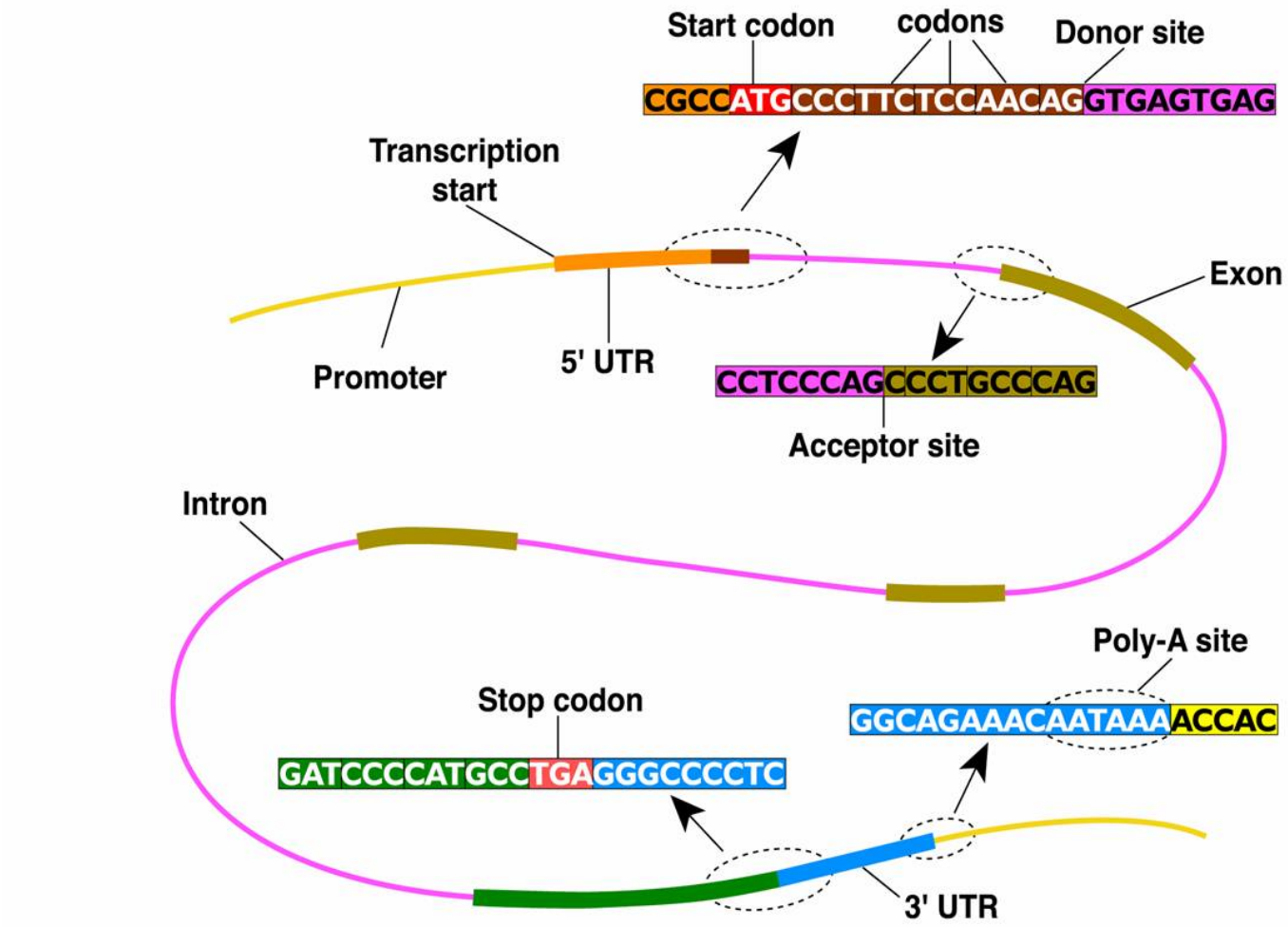
E.g., Predicate-Argument Structures



E.g., Alignments



Gene Finding and Analysis



Slide due to E. Xing

Phylogenetic Relationships

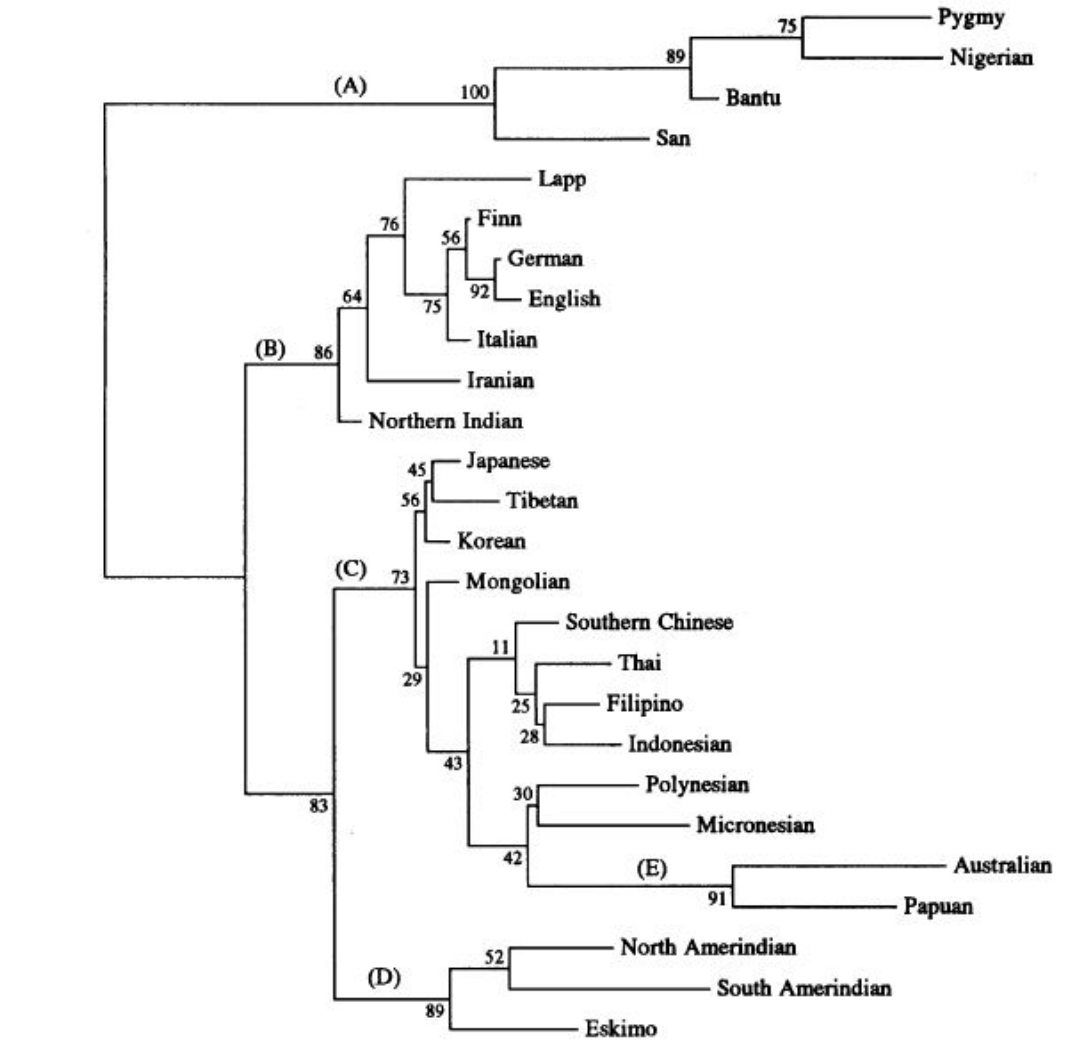
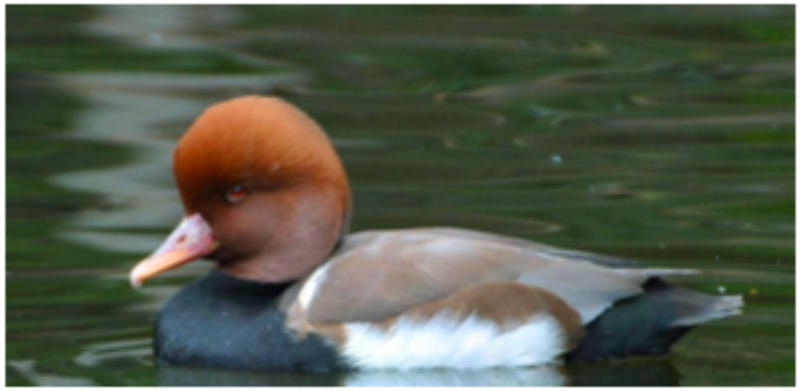


Image Segmentation



from Nowozin and Lampert (2010)

Implications of "Going Structured"

- All aspects of training and testing become more complex:
  - Designing a model
  - Prediction algorithms (once you have a model)
  - Learning your model from data
  - Measuring "error" of a prediction
- Machine learning helps with "mental hygiene"!
  - Principles that will help you explain and understand your methods
  - Generic optimization algorithms
  - Formal guarantees (sometimes)
  - Baselines when you're tackling a new problem

The Structured Prediction Way

1. Formally define the inputs and outputs.
2. Identify a scoring function over input-output pairs, and an algorithm that can find the maximum-scoring output given an input.
3. Determine what data can be used to learn to predict outputs from inputs, and apply a learning algorithm to tune the parameters of the scoring function.
4. Evaluate the model on an objective criterion measured on unseen test data.

Topics

- Inference (ch. 2, 5)
- Learning from Complete Data (ch. 3)
- Learning from Incomplete Data (ch. 4)

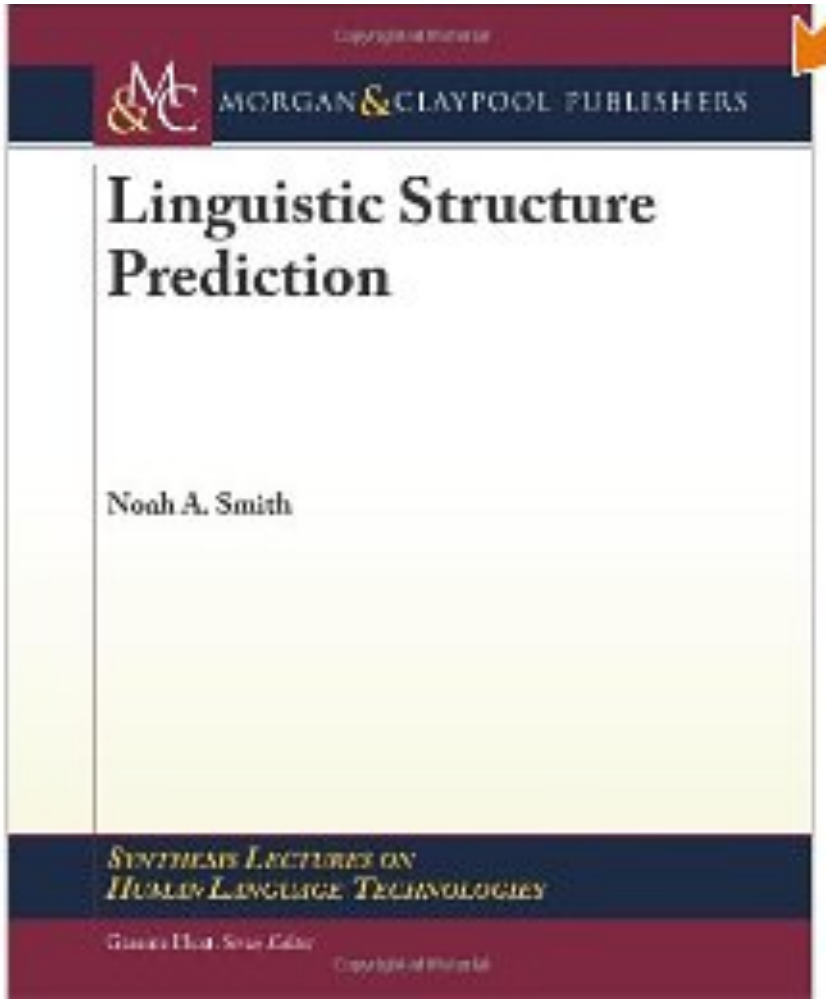
Format of the Course

- About five assignments (12 points each)
- Survey paper
  - 20 points spread over the term
  - 20 points for the final paper
- No exams

Email list:

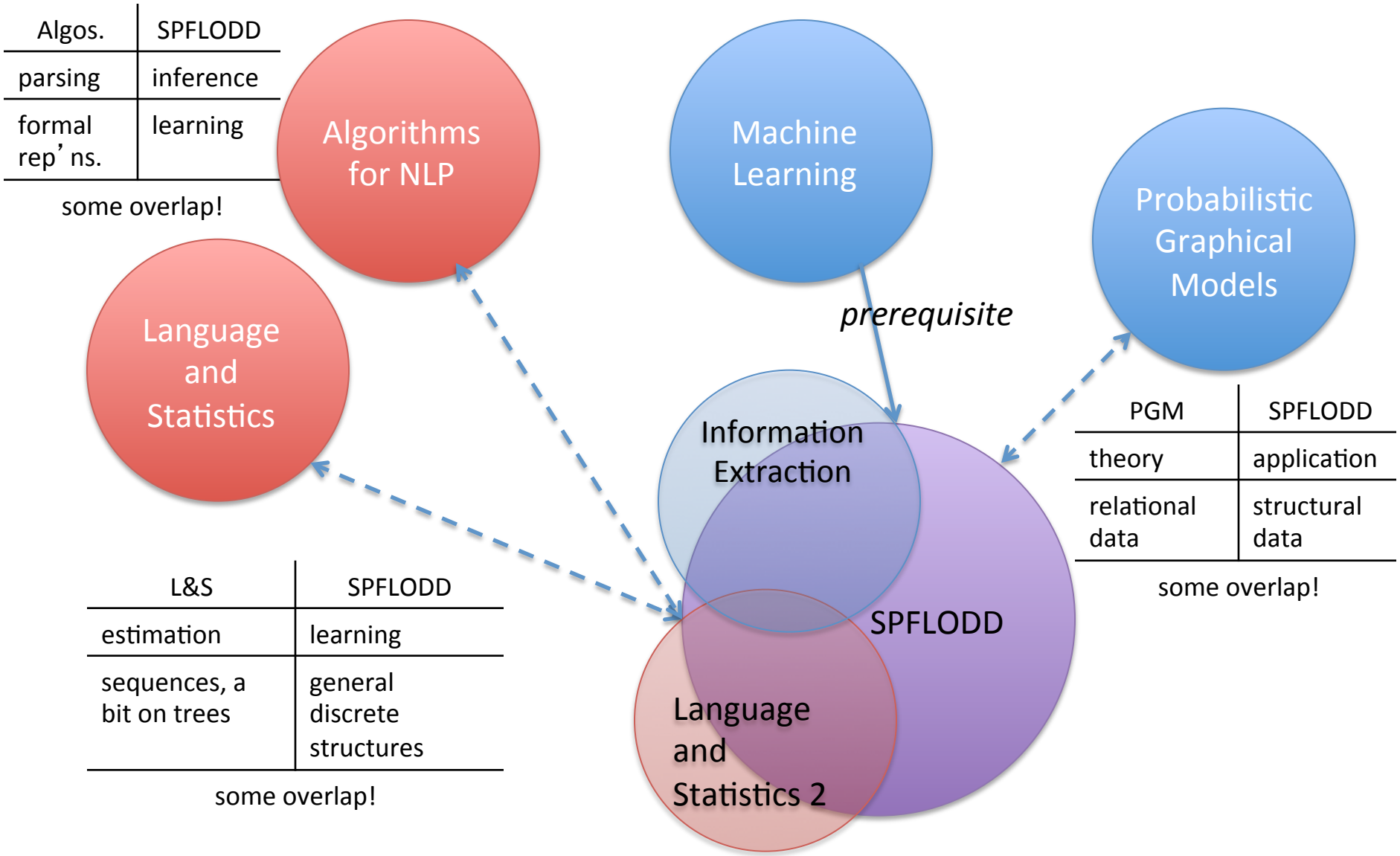
https://mailman.srv.cs.cmu.edu/mailman/listinfo/11763-announce

The Book



- *Linguistic Structure Prediction*
- Available in electronic form (free at CMU) and print form.

SPFLODD and Other Classes



Homework for Thursday

- Read *LSP*, preface and chapter 1.