Introduction and language models

CS 690N, Spring 2018

Advanced Natural Language Processing

http://people.cs.umass.edu/~brenocon/anlp2018/

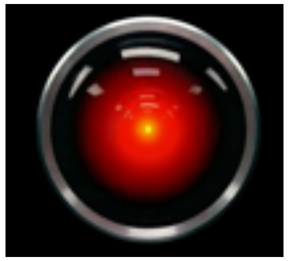
Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

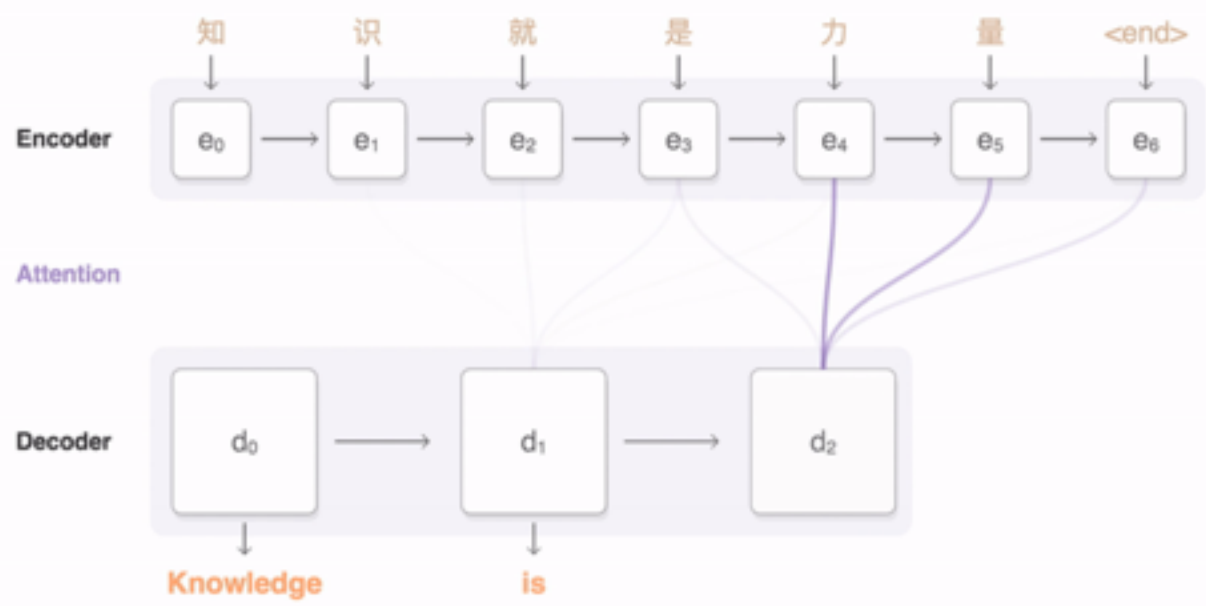
[including slides from Andrew McCallum and Chris Manning/Dan Jurafsky]

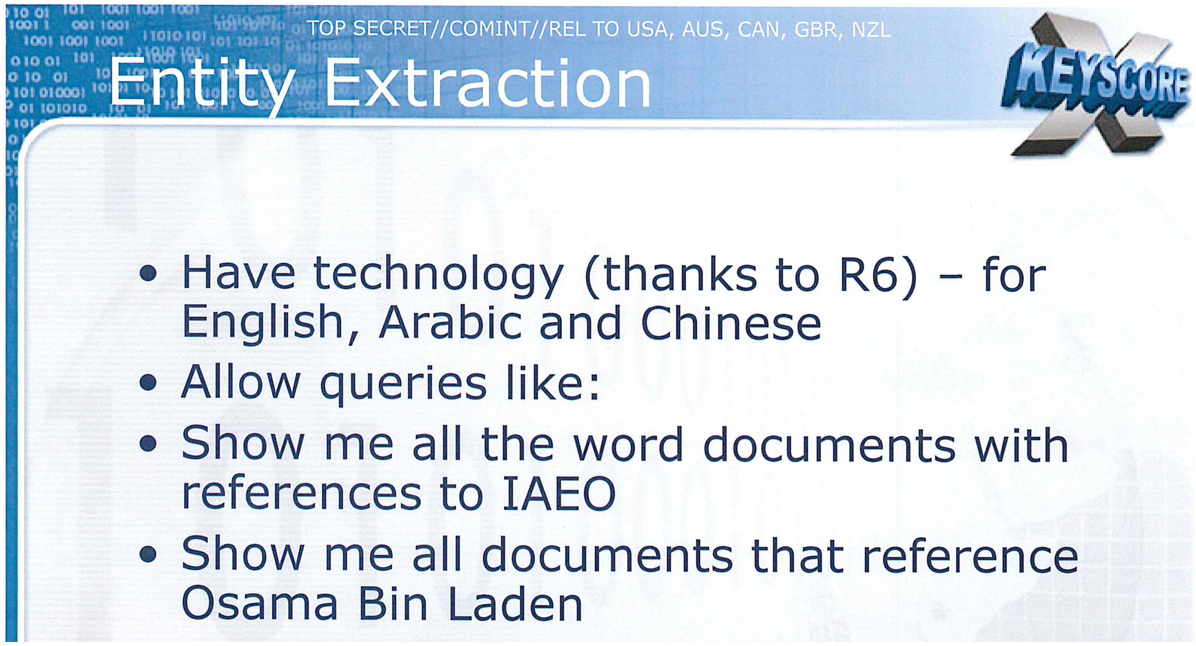
Computation + Language



What can I help you with?

Google Research Blog





TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

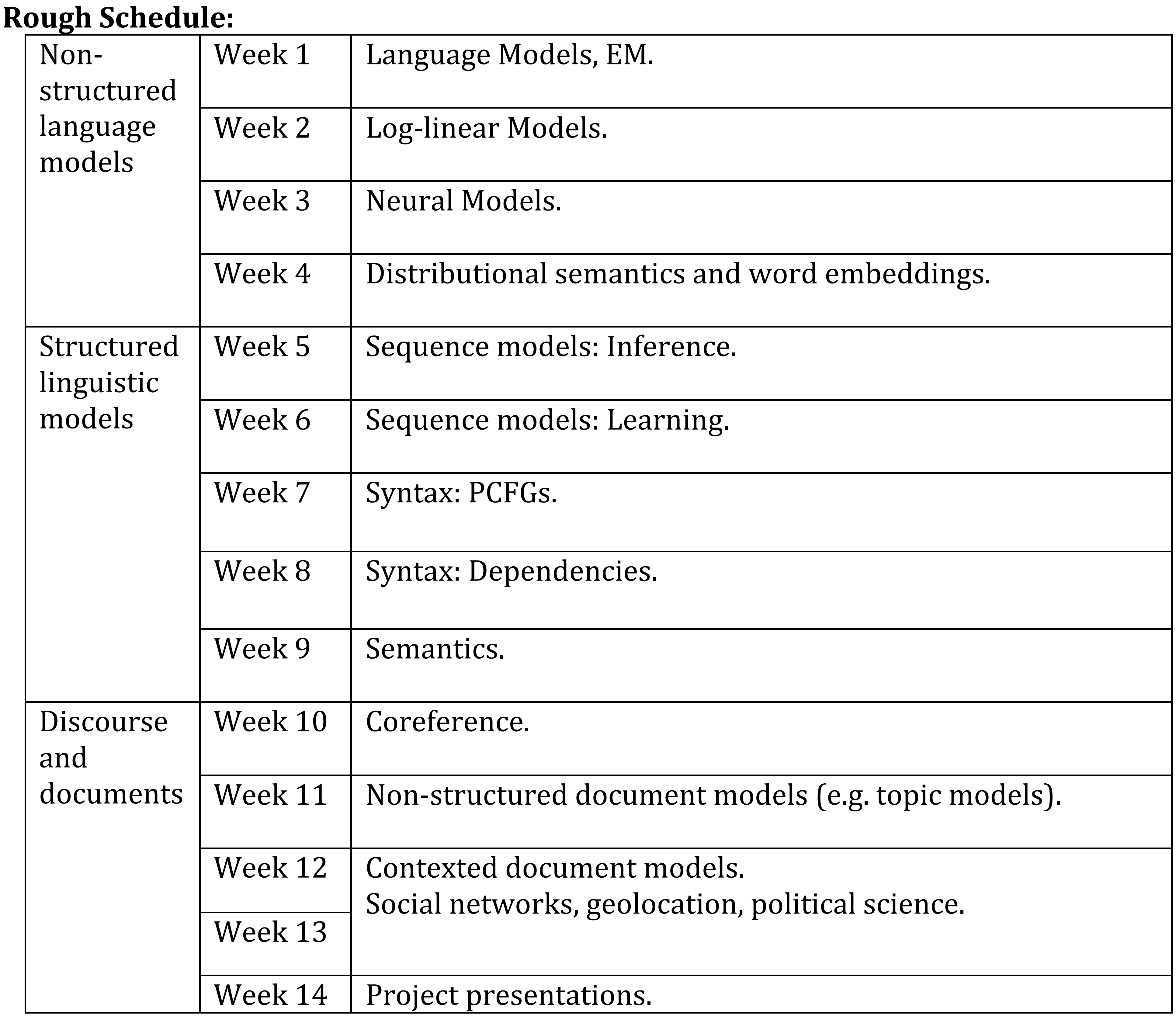
Entity Extraction

KEYSCORE

- Have technology (thanks to R6) – for English, Arabic and Chinese
- Allow queries like:
- Show me all the word documents with references to IAEO
- Show me all documents that reference Osama Bin Laden

- Learn methods, models, and research scholarship in natural language processing
  - Goal: be able to read, and ideally produce, current NLP research at ACL, EMNLP, NIPS, etc.
- Course components
  - Homeworks (30%, likely 3 total) -- programming, experiments, writing. First will be released this week.
  - Paper summaries (10%) -- approx. weekly. One paragraph summarizing and reacting to an assigned paper. First is due in one week.
  - Research presentation (10%) -- once in the semester, present a research paper for the class.
  - Literature review (25%) -- approx. mid-way in semester. Document reviewing an area of the NLP/CL research literature
  - Final Project (25%) -- proposal, final report, in-class presentation

Rough topics schedule



Rough Schedule:

Non-structured language models	Week 1	Language Models, EM.
	Week 2	Log-linear Models.
	Week 3	Neural Models.
	Week 4	Distributional semantics and word embeddings.
Structured linguistic models	Week 5	Sequence models: Inference.
	Week 6	Sequence models: Learning.
	Week 7	Syntax: PCFGs.
	Week 8	Syntax: Dependencies.
	Week 9	Semantics.
Discourse and documents	Week 10	Coreference.
	Week 11	Non-structured document models (e.g. topic models).
	Week 12	Contexted document models. Social networks, geolocation, political science.
	Week 13	
	Week 14	Project presentations.

- Feedback #0 is on Moodle, due tomorrow -- **why are you taking this course?**
- Feedback #1 is due in one week (next Tuesday) on the Rosenfeld (1996) reading

Your TA: Katie Keith

Language is hard (ambiguity)

Language is hard (ambiguity)

- Juvenile Court to Try Shooting Defendant

Language is hard (ambiguity)

- Juvenile Court to Try Shooting Defendant
- Hospitals Are Sued by 7 Foot Doctors

Language is hard (ambiguity)

- Juvenile Court to Try Shooting Defendant
- Hospitals Are Sued by 7 Foot Doctors
- Alice saw Bob with a telescope.

Language is hard (ambiguity)

- Juvenile Court to Try Shooting Defendant
- Hospitals Are Sued by 7 Foot Doctors
- Alice saw Bob with a telescope.
- Our company is training workers.

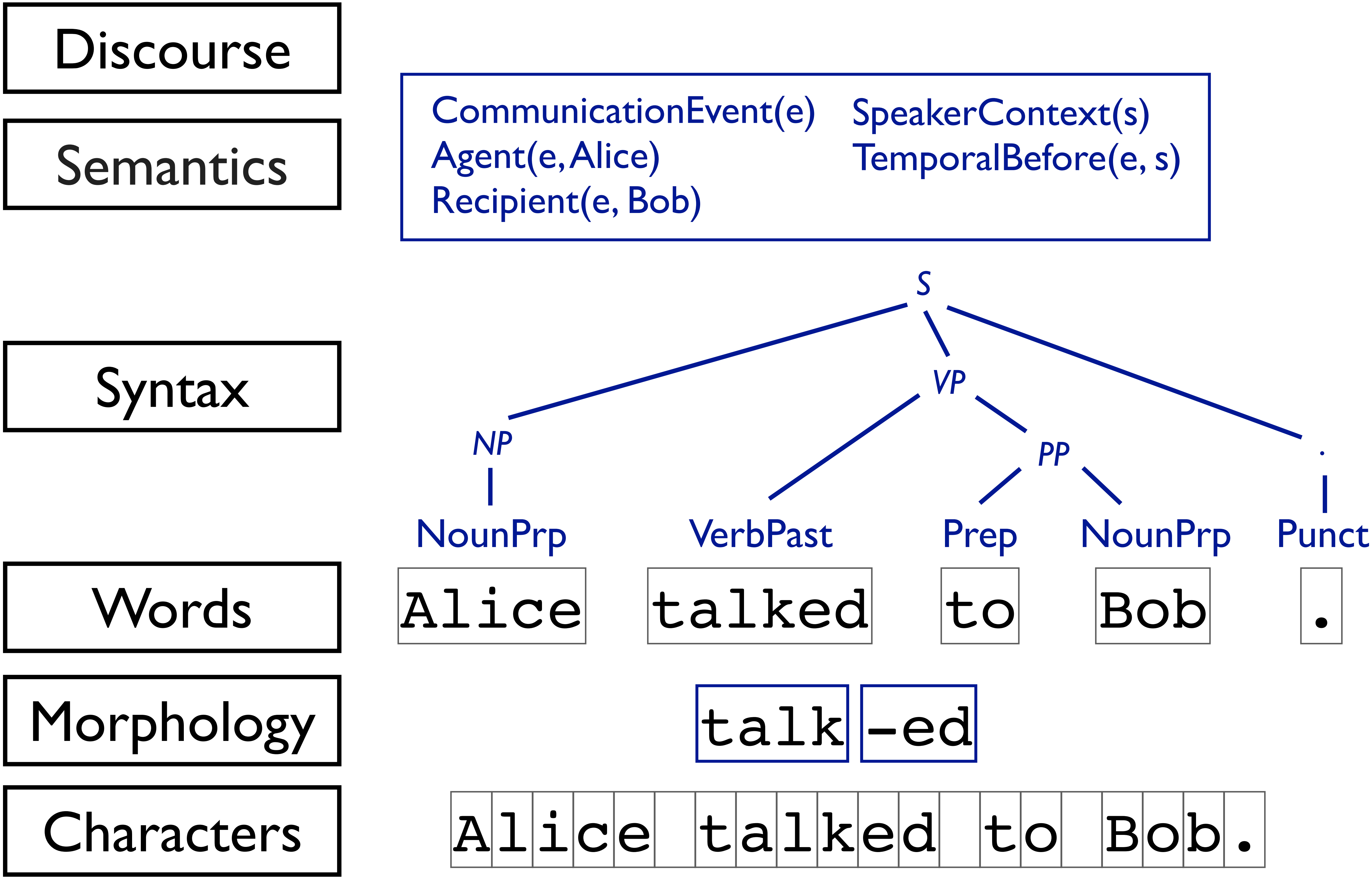
Language is hard (ambiguity)

- Juvenile Court to Try Shooting Defendant
- Hospitals Are Sued by 7 Foot Doctors
- Alice saw Bob with a telescope.
- Our company is training workers.
- They found that in order to attract settlers -- and make a profit from their holdings -- they had to offer people farms, not just tenancy on manorial estates.

What should NLP do?

- What would full natural language understanding mean?
- Contrast?: Typical NLP tasks
  - Text classification
  - Recognizing speech
  - Web search
  - Machine translation
  - Part-of-speech tagging

Levels of linguistic structure



Discourse

Semantics

Syntax

Words

Morphology

Characters

CommunicationEvent(e)
Agent(e, Alice)
Recipient(e, Bob)

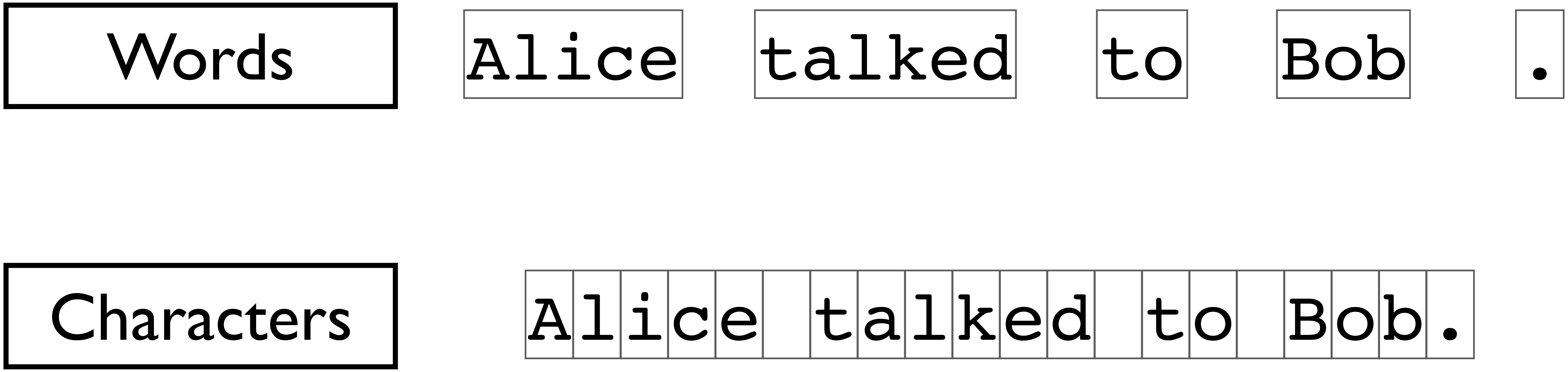
SpeakerContext(s)
TemporalBefore(e, s)

Alice talked to Bob .

talk -ed

A l i c e   t a l k e d   t o   B o b .

Levels of linguistic structure



Words

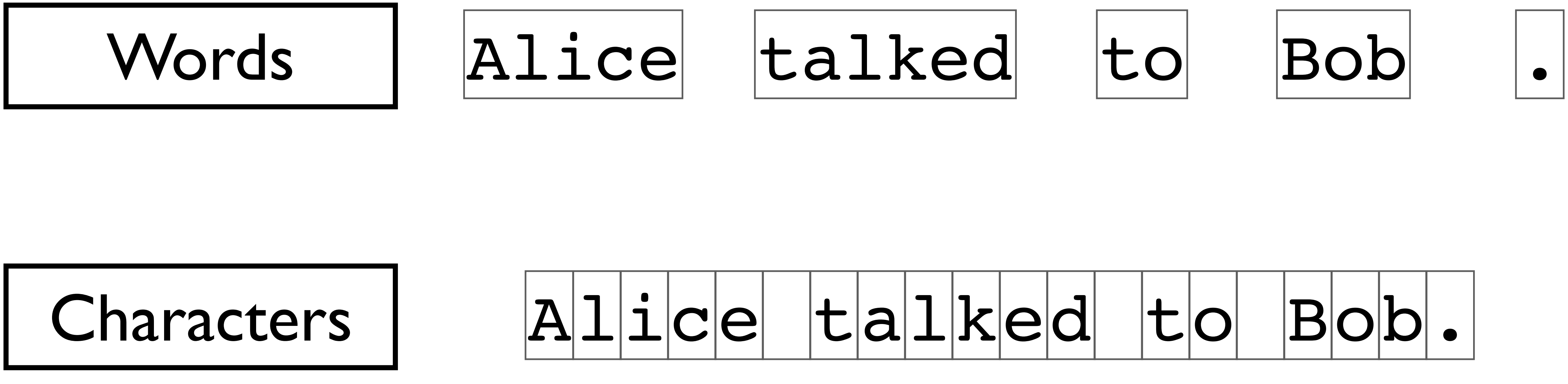
Alice talked to Bob .

Characters

Alice talked to Bob.

Levels of linguistic structure

Words are fundamental units of meaning



Words

`Alice` `talked` `to` `Bob` `.`

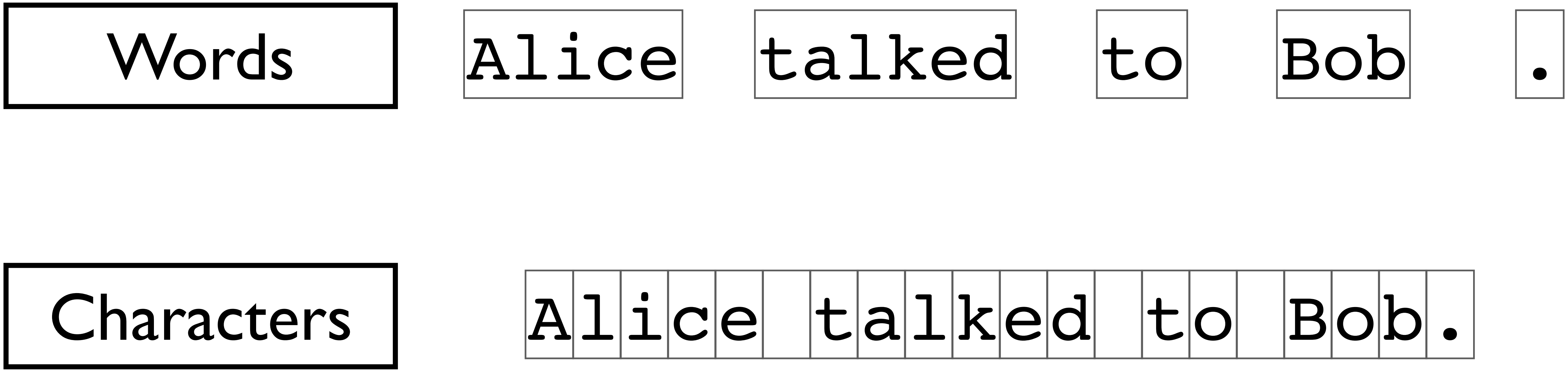
Characters

`A` `l` `i` `c` `e` ` ` `t` `a` `l` `k` `e` `d` ` ` `t` `o` ` ` `B` `o` `b` `.`

Levels of linguistic structure

Words are fundamental units of meaning
and easily identifiable*

*in some languages
*depending what you mean by "word"



Words

Alice talked to Bob .

Characters

Alice talked to Bob.

Modeling paradigms

- Discriminative vs. Generative
  - Linguistic structured prediction
    P(Structure | Text)
    - Structure = word categories, entity recognition, semantic relations, etc.
    - Discriminative modeling often used
  - Generative modeling
    P(Text)
    - Could use latent structure model
      = $\sum_{\text{structure}}$ P(Text, Structure)
      - Can use for structured prediction via Bayes Rule
    - Conditional generation (e.g. translation): P(Text1 | Text2)
    - Exploratory use: topic models
- Structure vs. Non-structured
  - Bag-of-words, sequences, trees, graphs
  - How to deal with combinatorial explosions? (graphical models, finite-state and stack automata, search heuristics...)

Language Models

Language Models

- P(text): Probability of generating a sequence of symbols

Language Models

- P(text): Probability of generating a sequence of symbols
    - High prob vs low prob sentences

Language Models

- P(text): Probability of generating a sequence of symbols
  - High prob vs low prob sentences
- Why?

Language Models

- P(text): Probability of generating a sequence of symbols
  - High prob vs low prob sentences
- Why?
  - Science: Explain humans' generative capacity for language

Language Models

- P(text): Probability of generating a sequence of symbols
  - High prob vs low prob sentences
- Why?
  - Science: Explain humans' generative capacity for language
  - Engineering: Fluency in language generation

Language Models

- Try to model just one sentence/utterance at a time
- Whole-sentence MLE?
- Problem: Learning from sparse data vs. generative capacity of language

The Chain Rule applied to compute joint probability of words in sentence

$$P(w_1 w_2 \dots w_n) = \prod_i P(w_i \mid w_1 w_2 \dots w_{i-1})$$

P("its water is so transparent") =

P(its) × P(water|its) × P(is|its water)

× P(so|its water is) × P(transparent|its water is so)

Markov chain models

- **Markov process**: words are generated one at a time. Process ends when END symbol is emitted.
- **First-order Markov assumption**: Assume a word depends only on previous word

$$P(w_t|w_1..w_{t-1}) = P(w_t|w_{t-1})$$

- This yields joint probability

$$P(w_1..w_T) = \prod_t P(w_t \mid w_1..w_{t-1})$$ <-- chain rule

$$= \prod_t P(w_t \mid w_{t-1})$$ <-- Markov assumption

Markov (1913)

1856 - 1922

- Took 20,000 characters from Pushkin's *Eugene Onegin* to see if it could be approximated by a first-order chain of characters.

vowel	consonant
0.43	0.57

0th order model

	c_t = vowel	c_t = consonant
c_{t-1} = vowel	0.13	0.87
c_{t-1} = consonant	0.66	0.34

1st order model

Andrew McCallum, UMass Amherst, Slide material from Dan Klein

Markov Approximations to English

- Zero-order approximation, P(c)
  - XFOML RXKXRJFFUJ ZLPWCFWKCRJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD
- First-order approximation, P(c|c)
  - OCRO HLI RGWR NWIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA
- Second-order approximation, P(c|c,c)
  - ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE

[Shannon 1948]

Big Data is still not infinite

Noam Chomsky (*Syntactic Structures*, 1957)
Responding to Markov & Shannon -type approaches

Sentences (1) and (2) are equally nonsensical, but any speaker of English will recognize that only the former is grammatical.

(1) Colorless green ideas sleep furiously.
(2) Furiously sleep ideas green colorless.

[T]he notion "grammatical in English" cannot be identified in any way with the notion "high order of statistical approximation to English". It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English.

Dealing with data sparsity

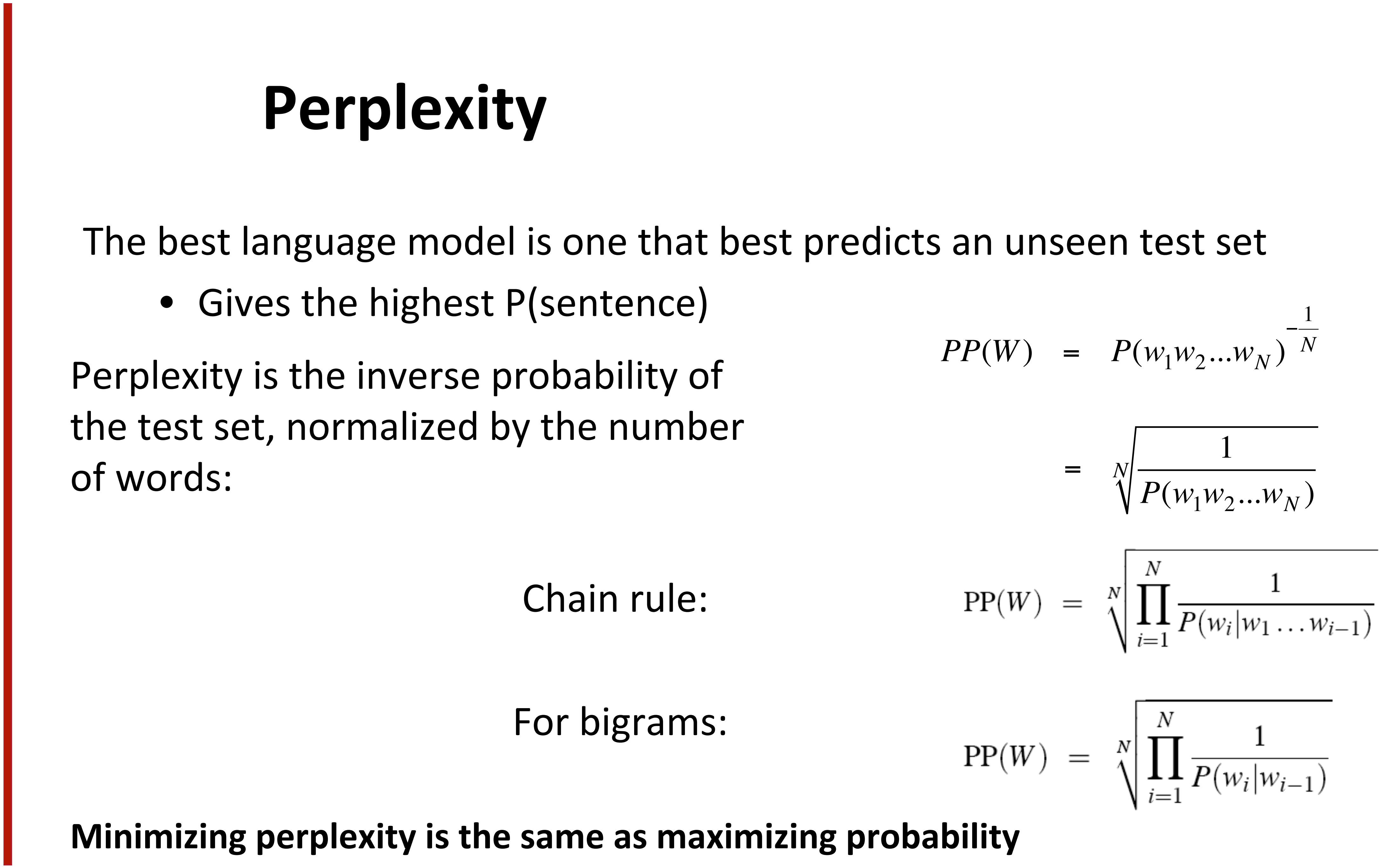
- Within n-gram models
  - What's wrong with MLE?
  - Backoff and interpolation:
    combine different Markov orders
  - Smoothing (pseudocounts, discounting):
    observed data counts for less

- Latent/hidden variables
  - Linguistic structure
  - Generalizable word attributes?
  - Long-distance dependencies?

Evaluation

- Does the LM prefer good sentences to bad ones?
- Extrinsic vs. Intrinsic eval

- Typical proxy task: held-out likelihood/perplexity
    - Does the LM give high probability to real text from a test set?



Perplexity

The best language model is one that best predicts an unseen test set

- Gives the highest P(sentence)

Perplexity is the inverse probability of the test set, normalized by the number of words:

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}}$$

Chain rule:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1 \ldots w_{i-1})}}$$

For bigrams:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}}$$

Minimizing perplexity is the same as maximizing probability

- H(p): Entropy(p)
- H(p,q): Cross-Entropy(true p, predicted q)
- Perplexity = exp(CrossEntropy)