

# Introduction and language models

**CS 690N, Spring 2017**

Adv. Natural Language Processing

<http://people.cs.umass.edu/~brenocon/anlp2017/>

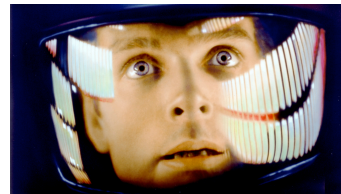
**Brendan O'Connor**

College of Information and Computer Sciences

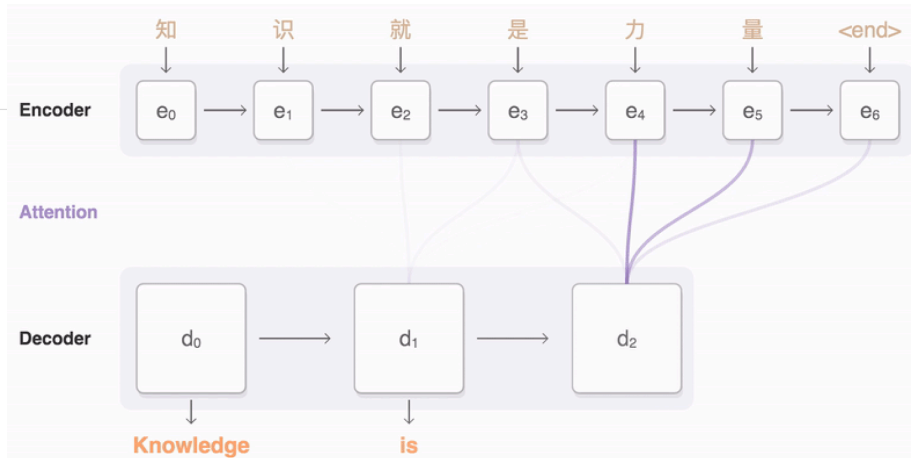
University of Massachusetts Amherst

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

# Computation + Language




Google Research Blog



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

## Entity Extraction



- 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 and models in natural language processing
  - Goal: be able to read, and ideally produce, current NLP research at ACL, EMNLP, NIPS, etc.
- Course components
  - Homeworks -- programming, experiments, writing
  - Project -- proposal, progress report, (poster?) presentation, final report

# Rough topics schedule

Non-structured language models	Week 1	Language Models. Information theory. Multinomials, EM. <i>[Assignment 1: Pereira 2000]</i>
	Week 2	Log-linear Models.
	Week 3	Neural Models. <i>[Assignment 2]</i>
	Week 4	LM Applications.
Structured (linguistic) analysis	Week 5	Rules. Information extraction, shallow parsing.
	Week 6	Syntax. PCFGs. <i>[Project proposal]</i>
	Week 7	Structured Prediction. Parameter learning, sequence tagging. <i>[Assignment 3]</i>
	Week 8	Syntax. Dependencies. Neural network parsing.
	Week 9	Semantics. Argument realization, Davidsonian representations, relation extraction. <i>[Assignment 4.]</i> <i>[Project milestone report.]</i>
Discourse and documents	Week 10	Coreference.
	Week 11	Non-structured document models. Topic models, log-linear BOW. <i>[Assignment 5]</i>
	Week 12	Contexted document models. Social networks, geolocation, political science.
	Week 13	
	Week 14	Project presentations. <i>[Project final report: end of finals]</i>

# 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
  - Part-of-speech tagging

# Levels of linguistic structure

Discourse

Semantics

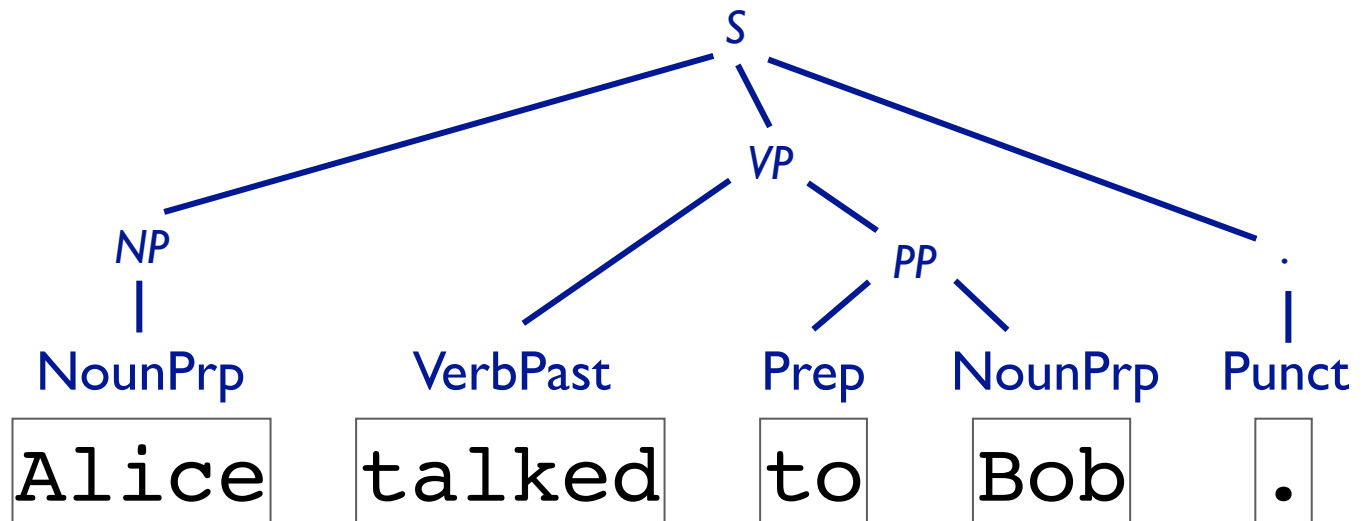
Syntax

Words

Morphology

Characters

CommunicationEvent(e) SpeakerContext(s)  
Agent(e, Alice) TemporalBefore(e, s)  
Recipient(e, Bob)



talk -ed

Alice talked to Bob.

# Levels of linguistic structure

Words are fundamental units of meaning  
and easily identifiable\*

\*in some languages

Words

Alice

talked

to

Bob

.

Characters

Alice talked to Bob.



# Language Models

- $P(\text{text})$ : 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

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

$P(\text{"its water is so transparent"}) =$

$P(\text{its}) \times P(\text{water} | \text{its}) \times P(\text{is} | \text{its water})$

$\times P(\text{so} | \text{its water is}) \times P(\text{transparent} | \text{its water is so})$

# 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

$$\begin{aligned} P(w_1 .. w_T) &= \prod_t P(w_t | w_1 .. w_{t-1}) && \leftarrow \text{chain rule} \\ &= \prod_t P(w_t | w_{t-1}) && \leftarrow \text{Markov assumption} \end{aligned}$$

# 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

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

1st order model

# 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 ILO NASIVE 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
  - 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  $P(\text{sentence})$

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

$$\begin{aligned} PP(W) &= P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}} \end{aligned}$$

Chain rule:

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

For bigrams:

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

**Minimizing perplexity is the same as maximizing probability**

[Board: LL and perplexity]