Language Modeling

Prof. Sameer Singh

CS 295: STATISTICAL NLP

WINTER 2017

January 24, 2017

Based on slides from Dan Jurafsky, Noah Smith, and everyone else they copied from.

Outline

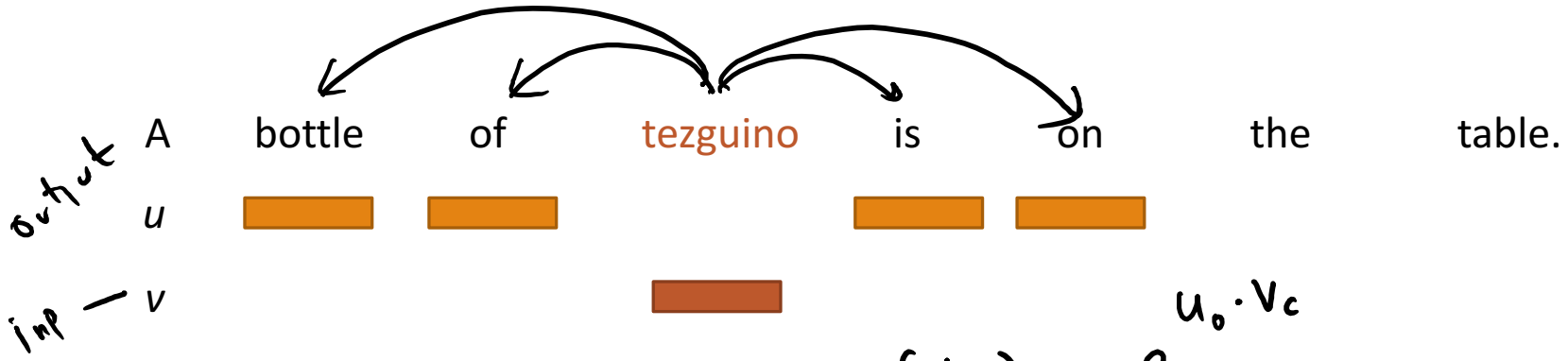
- Wrapup Word Embeddings
- Introduction to Language Models
- N-Gram Based Language Models
- Smoothing Language Models

Outline

- Wrapup Word Embeddings
- Introduction to Language Models
- N-Gram Based Language Models
- Smoothing Language Models

Predict surrounding words

$$P(w_{t+j} | w_t) \quad \forall\, j \in \{-m, \ldots m\},\ j \neq 0$$



$$P(o|c) = \frac{e^{u_o \cdot v_c}}{\sum_{w}^{V} e^{u_w \cdot v_c}}$$

Negative Sampling

$$\tilde{p}(o|c) = \frac{e^{u_o \cdot v_c}}{1 + e^{u_o \cdot v_c}} \quad \text{for } o = 0, \ldots, 1$$

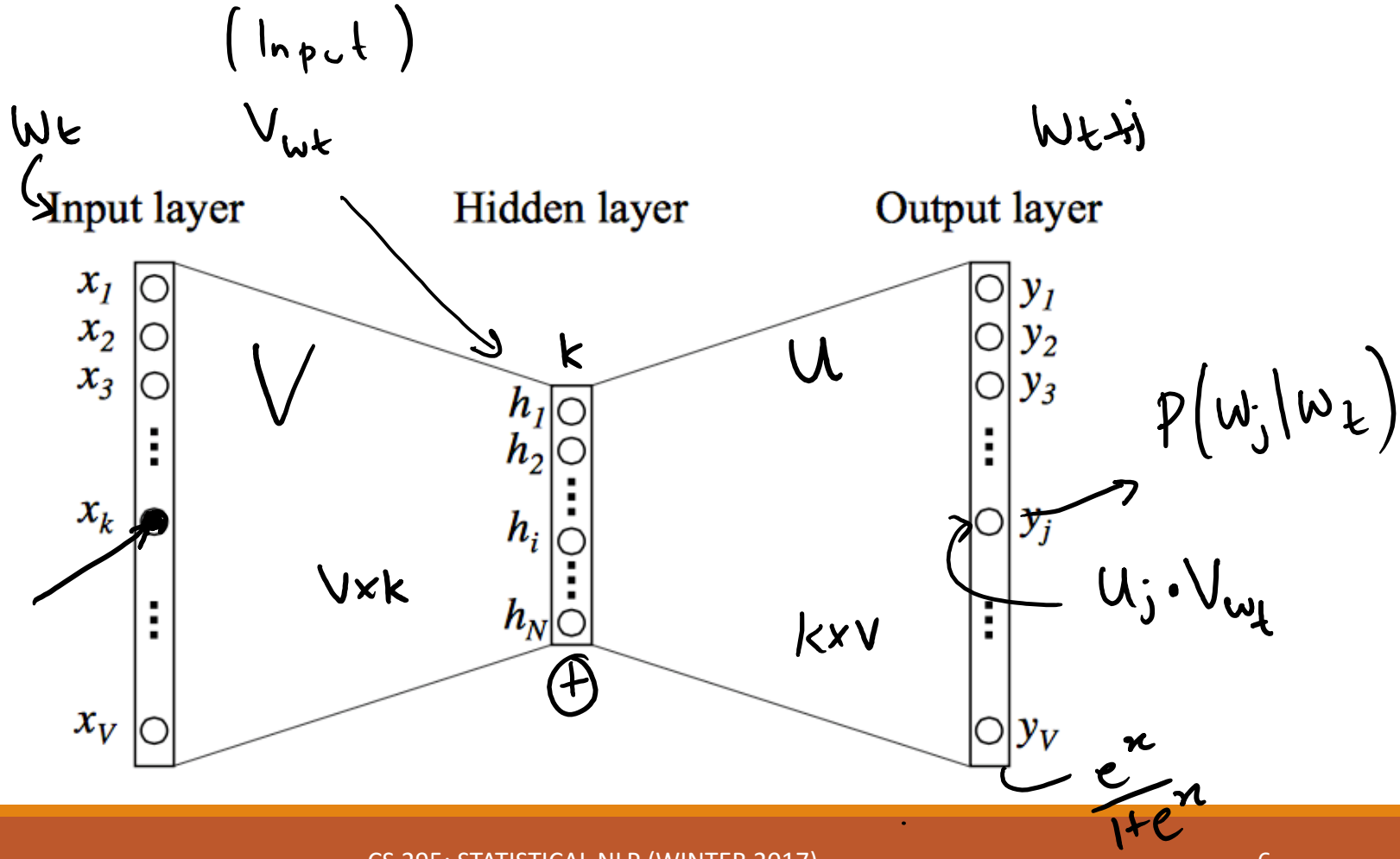
$$\sum_w \tilde{p}(w|c) \neq 1$$

$$\underset{V,U}{\text{argmax}} \sum_t \sum_j \log \tilde{p}(w_{t+j}|w_t) + \sum_k \frac{1}{k} \log\left(1 - \tilde{p}(w_k|w_t)\right)$$

Negative Sampling

weeks, days ⟶ minutes

Neural View of Embeddings



Word embeddings

Variations

- Skip-gram: predict context from word
- CBOW: predict word from context bag of words
- Dependencies: a better description of context

Uses

- Similarity: $w_i, w_j \sim \cos(v_i, v_j)$
- Grammar: walking - walk + swimming → swam
- Analogies
  - Gender: King - male + female → queen
  - Facts: Doctor - m + f → nurse; Capital - Country + France → Paris

Outline

Wrapup Word Embeddings

Introduction to Language Models

N-Gram Based Language Models

Smoothing Language Models

Language Models

Probability of a Sentence

$$P(W) = P(w_1 w_2 \ldots w_n)$$

- Is a given sentence something you would expect to see?
- Syntactically (grammar) and Semantically (meaning)

Probability of the Next Word

$$P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

$$\in \{1 \ldots V\}$$

- Predict what comes next for a given sequence of words.
- Think of it as V-way classification

Task: Speech Recognition

"eyes awe of an"

OR

"I saw a van"

word sequence	$\log p(\text{acoustics} \mid \text{word sequence})$
the station signs are in deep in english	-14732
the stations signs are in deep in english	-14735
the station signs are in deep into english	-14739
the station 's signs are in deep in english	-14740
the station signs are in deep in the english	-14741
the station signs are indeed in english	-14757
the station 's signs are indeed in english	-14760
the station signs are indians in english	-14790
the station signs are indian in english	-14799
the stations signs are indians in english	-14807
the stations signs are indians and english	-14815

Task: Machine Translation

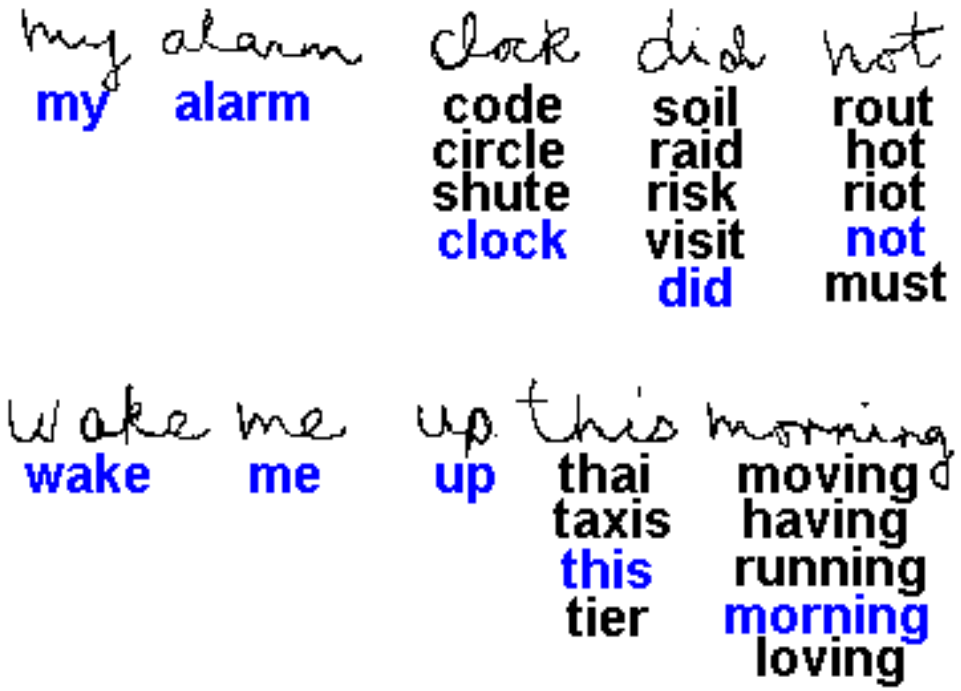
Quiero ir a la playa más bonita.

I try | to leave | per | the most lovely | open space.

I want | to go | to | the prettiest | beach.

$P(\vec{w}_2) > P(\vec{w}_1)$

Task: Handwriting Recognition



http://www.cedar.buffalo.edu/handwriting/HRoverview.html

Task: Image Captioning

A person skiing down a snow covered slope.

Task: Spelling Correction

The office is about fifteen minuets from my house

P(about fifteen minutes from) >> P(about fifteen minuets from)

Other Applications

Summarization

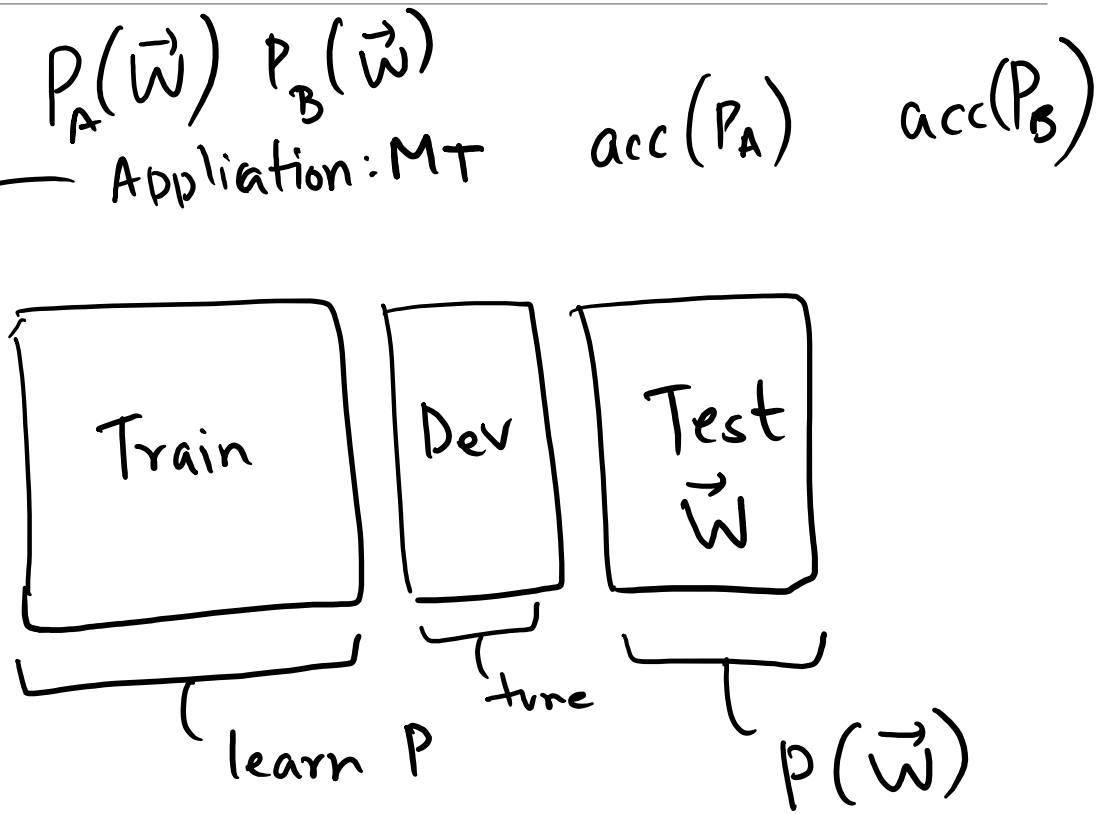
Question Answering

Dialog Systems

Evaluating Language Models

Best choice: Extrinsic

2nd choice: Intrinsic



Perplexity, PP

$$P(W) = \prod_i P(\vec{w}_i)$$

$$\frac{1}{N} \log_2 P(W) = \frac{1}{N} \sum_i \log_2 P(\vec{w}_i)$$

$$PP(W) = 2^{-\frac{1}{N} \sum_i \log P(\vec{w}_i)}$$

$$= \sqrt[N]{\frac{1}{\prod_i P(\vec{w}_i)}}$$

Random : $P(w_i) \cdots) = \frac{1}{V}$ $PP(W) = V$

Perfectly : $PP(w) = 1$

Generating Text from an LM

```
s = [ ]  # prefix

do
    w ~ P(w|s)
    s += w

while w != "EOS" or maxLength
```

Outline

Wrapup Word Embeddings

Introduction to Language Models

N-Gram Based Language Models

Smoothing Language Models

Direct Language Modeling

P("I do not like green eggs and ham") = $\frac{\#(\text{"I do not lik...."})}{N}$ → number of sentences

P(w | "I do not like green eggs and ") = $\frac{\#(\text{"I do not .."} + w)}{\#(\text{"I donot .. and"})}$

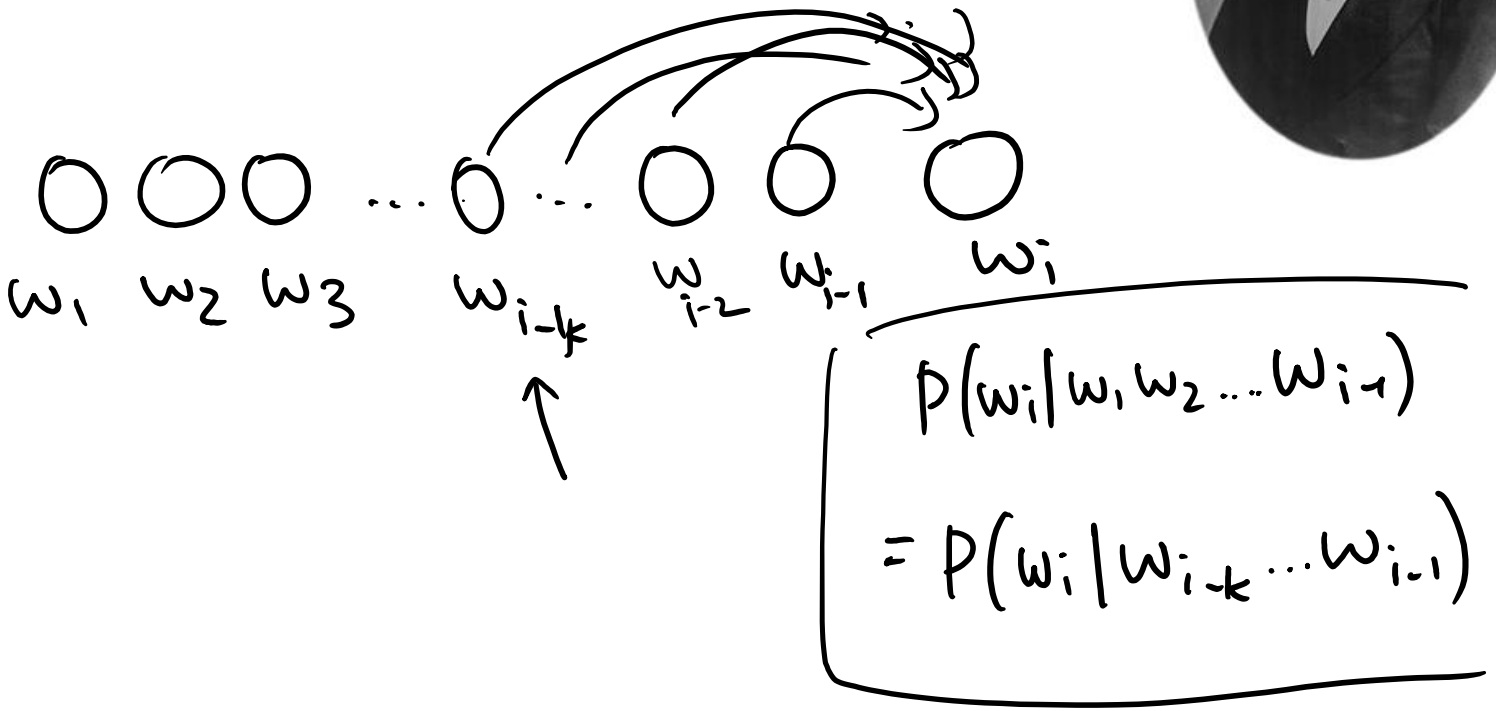
ham 1
w 0

Applying the Chain Rule

$$P(w_1 w_2 \ldots w_n) = P(w_1) P(w_2 \mid w_1) P(w_3 \mid w_1 w_2) \ldots$$
$$P(w_n \mid w_1 w_2 \ldots w_{n-1})$$

$$P(\text{"I do not like eggs"}) = P(\text{"I"} \mid \langle s \rangle)$$
$$P(\text{"do"} \mid \text{"I"})$$
$$\vdots$$
$$P(\text{"eggs"} \mid \text{"I do not like"})$$

Markov Assumption



Unigram Language Model

0th order Markov

$$P(w_i \mid w_1 \ldots w_{i-1}) = P(w_i)$$

$$= \frac{\#w_i}{N}$$

$N \rightarrow$ number of words

$$P(\text{"the a an is the"}) > P(\text{"I love food"})$$

Bigram Language Model

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

$$= \frac{\#\text{“}w_{i-1}\ w\text{”}}{\#\text{“}w_{i-1}\text{”}}$$

Corpus : 800k
└ Vocab 30k

30k x 30k

300k bigrams obs.

Berkeley Restaurant Project

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Berkeley Restaurant Project

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

N-Gram Language Models

$$P(w_i | w_1 \ldots w_{i-1}) = P(w_i | w_{i-n} \ldots w_{i-1})$$

$n = 3$ trigram

$= 4$ Quadgram

$\vdots$

"The computer which I had just put into the dining room on the fifth floor crashed."

"The computer which I had just put into the dining room on the fifth floor had lunch."

Shakespeare

Unigram

To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have
Every enter now severally so, let
Hill he late speaks; or! a more to leg less first you enter
Are where exeunt and sighs have rise excellency took of.. Sleep knave we. near; vile like

Bigram

What means, sir. I confess she? then all sorts, he is trim, captain.
Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.
What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman?

Trigram

Sweet prince, Falstaff shall die. Harry of Monmouth's grave.
This shall forbid it should be branded, if renown made it empty.
Indeed the duke; and had a very good friend.
Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.

Quadrigram

King Henry.What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in;
Will you not tell me who I am?
It cannot be but so.
Indeed the short and the long. Marry, 'tis a noble Lepidus.

Wall Street Journal

Unigram

Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

Bigram

Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

Trigram

They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

Implementation Tips

Use Logs

$\log P(w_i | w_{i-1} w_{i-2})$

- Prevent underflow
- Sums, instead of products

$\prod_i P(w_i) \Rightarrow \sum_i \log P(w_i)$

Filter out n-grams

- Rare n-grams are noisy/have low prob
- Use unigrams to filter bigrams…

count $> \gamma = 1, 2$

egg soup $> \gamma$

egg $> \gamma$

soup $> \gamma$

Outline

Wrapup Word Embeddings

Introduction to Language Models

N-Gram Based Language Models

Smoothing Language Models

Zero Probability Problem

Training set:

… denied the allegations
… denied the reports
… denied the claims
… denied the request

P("offer" | denied the) = 0

- Test set
  … denied the offer
  … denied the loan

Rare words/combinations

- Because corpus is finite..

Mispellings

- "minuets"

"UNK"

V

New words

- Truthiness
- #letalonethehashtags
- bigly

Laplace Smoothing

$$p(w_i \mid w_{i-1}) = \frac{\#(\text{“}w_{i-1}\, w_i\text{”}) + 1}{\#\text{“}w_{i-1}\text{”} + V}$$

Add-1 smoothing

$$\frac{\#(\text{“}w_{i-1}\, w_i\text{”}) + \lambda}{\#\text{“}w_{i-1}\text{”} + \lambda V}$$

Add λ smoothing

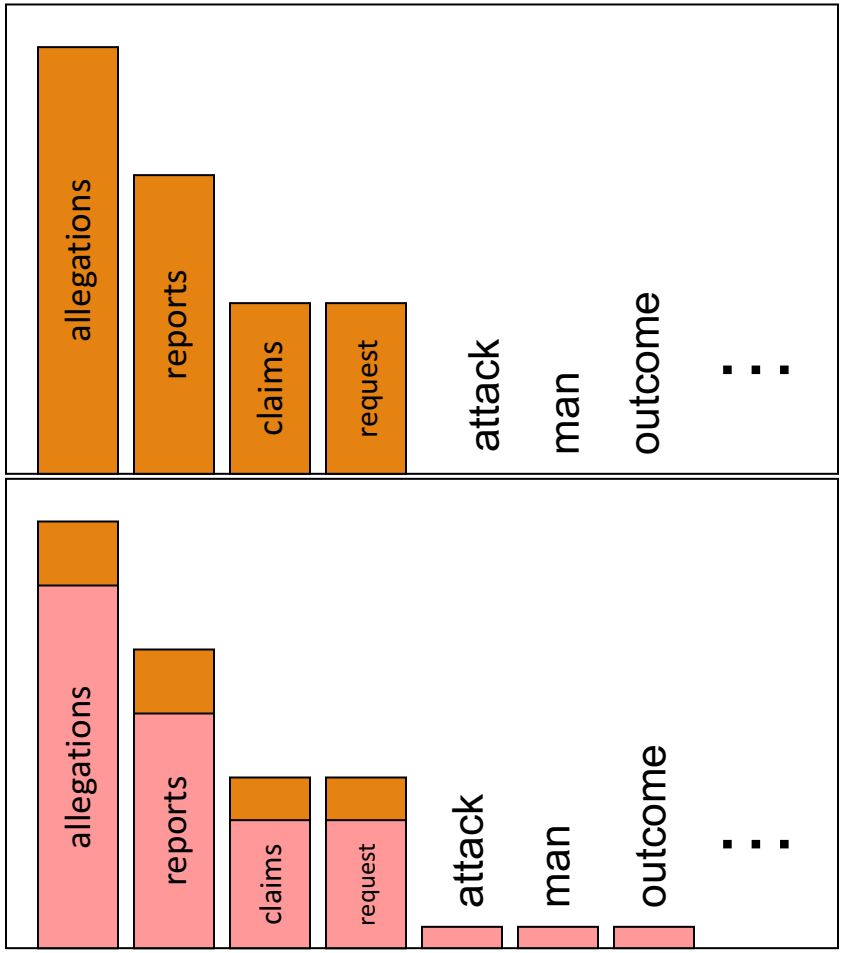
Intuition Behind Smoothing

When we have sparse statistics:

P(w | denied the)
3 allegations
2 reports
1 claims
1 request
7 total

Steal probability mass to generalize better

P(w | denied the)
2.5 allegations
1.5 reports
0.5 claims
0.5 request
2 other
7 total



Berkeley Restaurant Project

	i	want	to	eat	chinese	food	lunch	spend
i	6	828	1	10	1	1	1	3
want	3	1	609	2	7	7	6	2
to	3	1	5	687	3	1	7	212
eat	1	1	3	1	17	3	43	1
chinese	2	1	1	1	1	83	2	1
food	16	1	16	1	2	5	1	1
lunch	3	1	1	1	1	2	1	1
spend	2	1	2	1	1	1	1	1

Berkeley Restaurant Project

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

Backoff and Interpolation

Backoff

- Use trigram, unless rare
- Then use bigram, unless rare
- Then use unigram..

$$\hat{P}(w_i|w_{i-2}w_{i-1}) = \begin{cases} P(w_i|w_{i-2}w_{i-1}) & \#\text{"}w_{i-2}w_{i-1}w_i\text{"} > 0 \\ P(w_i|w_{i-1}) & \#\text{"}w_{i-1}w_i\text{"} > 0 \\ P(w_i) & \#\ w_i > 0 \\ \epsilon & \text{p.w.} \end{cases}$$

Interpolation

- Combine all three!
- Linear function with parameters
- Learn on held out data

$$\hat{P}(w_i|w_{i-2}w_{i-1}) = \lambda_1 P(w_i|w_{i-2}w_{i-1}) + \lambda_2 P(w_i|w_{i-1}) + \lambda_3 P(w_i)$$

$$\sum \lambda = 1 \quad \lambda_3 \leftarrow \text{context}$$

Upcoming...

Homework

- Homework 1 is due: January 26, 2017
- Write-up, data, and code for Homework 2 is up
- Homework 2 is due: February 9, 2017

Project

- Proposal is due: February 7, 2017 (~2 weeks)
- Make things more concrete: approach, metrics, baselines
- Mention progress, and address my concerns, if any
- Only 2 pages