CS 671 NLP LANGUAGE MODELING: N-GRAMS

amitabha mukerjee
iit kanpur

NLP Tasks

Word segmentation:

- Chinese: 浮法像蝴蝶.

  (“float like a butterfly)

- Hindi

  पांचफिरंगीअफसरोंकोफांसीपरलटकादिया

  - Q. Letter-or Syllable- based?
  - Which positions have low “sequence” probability?

NLP tasks and Probabilistic Models

- Other problems
  - Machine Translation:
    - P(**high** winds tonite) > P(**large** winds tonite)
  - Spell Correction
    - The office is about fifteen **minuets** from my house
      - P(about fifteen **minutes** from) > P(about fifteen **minuets** from)
  - Speech Recognition
    - P(I saw a van) >> P(eyes awe of an)
  - Verb argument structure discovery
    - Via factorization of syntactic parses to discover
    - Argument structure (syntax ?)
    - Selection preference (semantics)
  - + Summarization, question-answering, etc.,

Probabilistic Language Modeling

- Goal: compute the probability of a sentence or sequence of words:

  $P(W) = P(w_1,w_2,w_3,w_4,w_5 \ldots w_n)$

- Related task: probability of an upcoming word:

  $P(w_5|w_1,w_2,w_3,w_4)$

- A model that computes either of these:

  $P(W)$ or $P(w_n|w_1,w_2 \ldots w_{n-1})$ is called a **language model**.

- Better: **the grammar** But **language model** or **LM** is standard

Shannon Entropy

- Predict the next word/letter, given (*n-1*) previous letters or words : Fn = entropy = $\mathrm{SUM}_i (p_i \log p_i)$
- probabilities p_i (of n-grams) from corpus:
  - F_0 (only alphabet) = $\log_2 27$ = 4.76 bits per letter
  - F_1 (1-gram frequencies p_i) = 4.03 bits
  - F_2 (bigram frequencies) = 3.32 bits
  - F_3 (trigrams) = 3.1 bits
  - F_{word} = 2.62 bits
    (avg word entropy = 11.8 bits per 4.5 letter word)

Claude E. Shannon. "Prediction and Entropy of Printed English", *Bell System Technical Journal* 30:50-64. 1951.

Shannon Entropy : Human

- Ask human to guess the next letter:

```
THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG
----ROO------NOT-V-----I------SM----OBL---

READING LAMP ON THE DESK SHED GLOW ON
REA----------O------D----SHED-OLD--O-

POLISHED WOOD BUT LESS ON THE SHABBY RED CARPET
P-L-S-----O---BU--L-S—O-------SH-----RE—C------
```

- 69% guessed on 1st attempt [“-” = 1st attempt]

Claude E. Shannon. “Prediction and Entropy of Printed English”, *Bell System Technical Journal* 30:50-64. 1951.

Shannon Entropy : Human

- Count number of attempts:

```
T H E R E   I S   N O   R E V E R S E   O N   A   M O T O R C Y C L E   A
1 1 1 5 1 1 2 1 1 2 1 1 15 1 17 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1 1 1 3 1
F R I E N D   O F   M I N E   F O U N D   T H I S   O U T
8 6 1 3 1 1 1 1 1 1 1 1 1 1 1 6 2 1 1 1 1 1 1 2 1 1 1 1 1 1
R A T H E R   D R A M A T I C A L L Y   T H E   O T H E R   D A Y
4 1 1 1 1 1 1 1 1 5 1 1 1 1 1 1 1 1 1 1 1 6 1 1 1 1 1 1 1 1 1 1 1 1
```

- Entropy: F_1 =3.2, 4.0 F_{10} =1.0, 2.1 F_{100} = 0.6, **1.3**

Claude E. Shannon. “Prediction and Entropy of Printed English”, *Bell System Technical Journal* 30:50-64. 1951.

Language Modeling

- Examine short sequences of
  - letters
  - syllables
  - morphemes
  - words
- How likely is each sequence?
- **Markov Assumption** – word is affected only by its "prior local context" (last few words)

LANGUAGE MODELING

NL Corpora

Creating a Corpus

1961 : W. Nelson Francis and Henry Kucera of Brown Univ

500 samples of 2,000 words each from various text genres

→ American English

1970s : Lancaster-Oslo-Bergen corpus: British English

also 500 x 2000 = 1mn words – genres similar to Brown Corpus

Geoffrey Leech of Lancaster U.

1994: British National Corpus – 100mn words

Oxford U, Lancaster, Longman / Chambers dictionaries

10% : transcripts of spoken English

2000s: Google corpora: American english 155 bn words; British : 34bn

[Lindquist 2009]: Corpus linguistics and the description of English

The Brown Corpus

		# texts	%age
A	Press: reportage (newspapers)	44	8.8%
B	Press: editorial (including letters to the editor)	27	5.4%
C	Press: reviews (theatre, books, music, dance)	17	3.4%
D	Religion	17	3.4%
E	Skills and hobbies	36	7.2%
F	Popular lore	48	9.6%
G	Belles letters, biography, memoirs etc.	75	15.0%
H	Miscellaneous (mainly government documents)	30	6.0%
J	Learned (academic texts)	80	16.0%
K	General fiction (novels and short stories)	29	5.8%
L	Mystery and detective fiction	24	4.8%
M	Science fiction	6	1.2%
N	Adventure and Western fiction	29	5.8%
P	Romance and love story	29	5.8%
R	Humour	9	1.8%
	Non-fiction subtotal	374	75%
	Fiction subtotal	126	25%
	Total	500	100%

News: political, sports, society "spot news", financial, cultural)

[Lindquist 2009]: Corpus linguistics and the description of English

Parallel Corpora

Parallel Corpus

Congress MP from Haryana Birender Singh said at a programme that "once someone had told me that Rs 100 crore was required to get a Rajya Sabha berth.
But he said he got it for Rs 80 crore and saved Rs 20 crore. Now will people who are willing to invest Rs 100 crore, ever think of the poor country."

राज्य सभा सांसद बीरेंद्र सिंह ने एक कार्यक्रम में कहा था, "एक बार की बात है कि मुझे एक व्यक्ति ने बताया कि राज्य सभा की सीट 100 करोड़ रुपए में मिलती है. उसने बताया कि उसे खुद यह सीट 80 करोड़ रुपए में मिल गई, 20 करोड़ बच गए. मगर क्या वे लोग, जो 100 करोड़ खर्च करके यह सीट खरीदने के इच्छुक हैं, कभी इस गरीब देश के बारे में भी सोचेंगे?"

একটি অনুষ্ঠানে তিনি বলেন, 'আমাকে একজন বলেছিলেন, ১০০ কোটি রুপি হলেই রাজ্য সভার একটি আসন পাওয়া যায়।
তবে ৮০ কোটি রুপি দিয়ে তিনি একটি আসন সংগ্রহ করে ২০ কোটি রুপি বাঁচিয়েছেন।'

Matching on parallel Corpus

电脑**坏了。**
The computer is broken.
电脑**死机了。**
My computer has frozen.
我想玩电脑。
I want to play on the computer.
我家没有电脑。
I don't have a computer at home.
我有一台电脑。
I have a computer.
你有两台电脑吗？
Do you have two computers?

Parallel Corpus

电脑坏了。
The computer is broken.
电脑死机了。
My computer has frozen.
我想玩电脑。
I want to play on the computer.
我家没有电脑。
I don't have a computer at home.
我有一台电脑。
I have a computer.
你有两台电脑吗？
Do you have two computers?

电脑 : *diànnǎo*, computer
[电 : *diàn* lightning, electricity 脑 : *nǎo* brain]

Parallel Corpus

电脑**坏了。**
The computer is broken.
电脑**死机了。**
My computer has frozen.
我想玩电脑。
I want to play on the computer.
我家没有电脑。
I don't have a computer at home.
我有一台电脑。
I have a computer.
你有两台电脑吗？
Do you have two computers?

有 : “in possession of”
[**又** (“hand”) + 月 (**肉**) (“meat”) = a hand holding meat]

LANGUAGE MODELING

Generalization and zeros

The perils of overfitting

- N-grams only work well for word prediction if the test corpus looks like the training corpus
  - In real life, it often doesn’t
  - We need to train robust models that generalize!
  - One kind of generalization: Zeros!
    - Things that don’t ever occur in the training set
      - But occur in the test set

Zeros

- Training set:

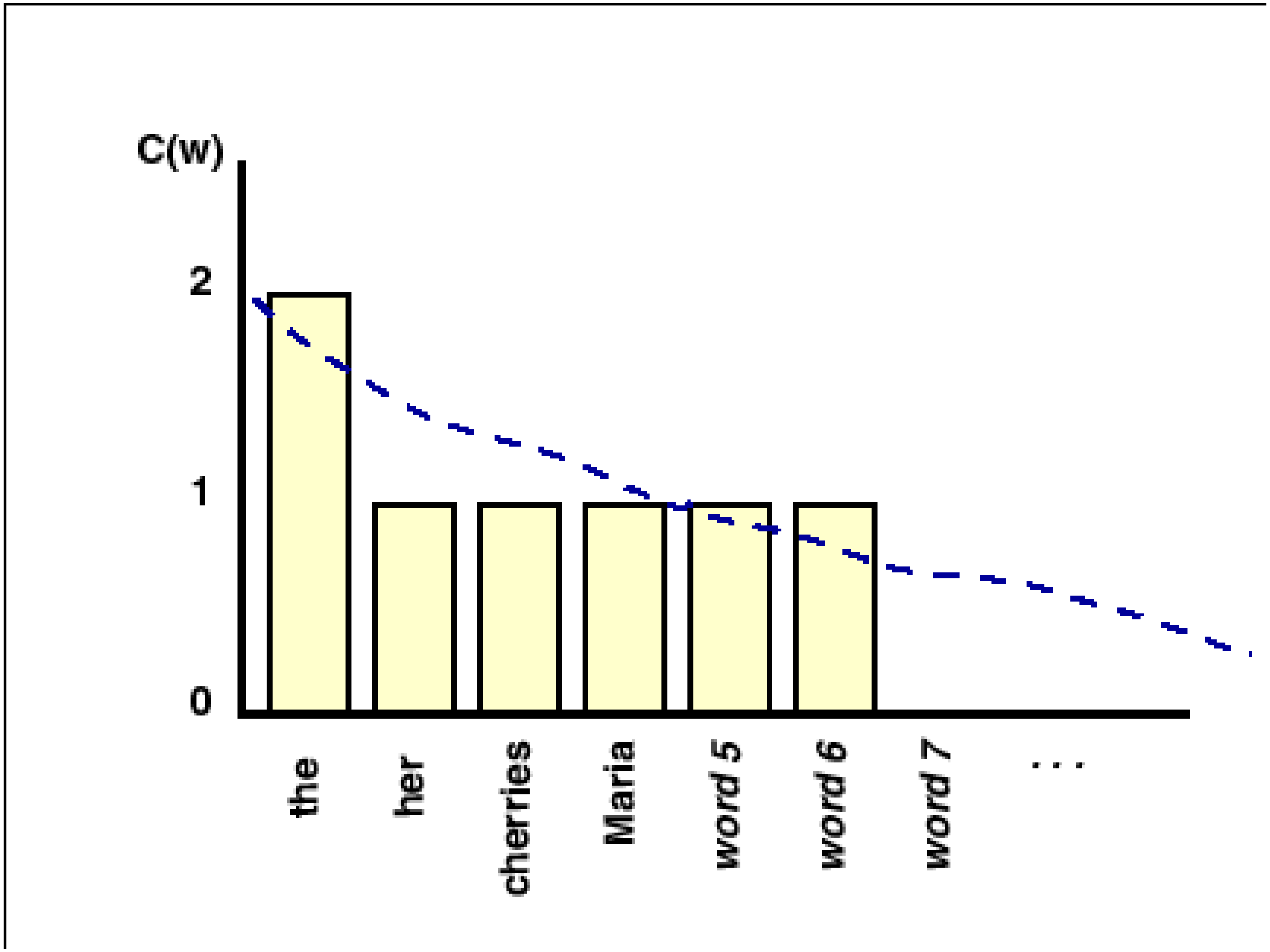
  … denied the allegations
  … denied the reports
  … denied the claims
  … denied the request

  P("offer" | denied the) = 0

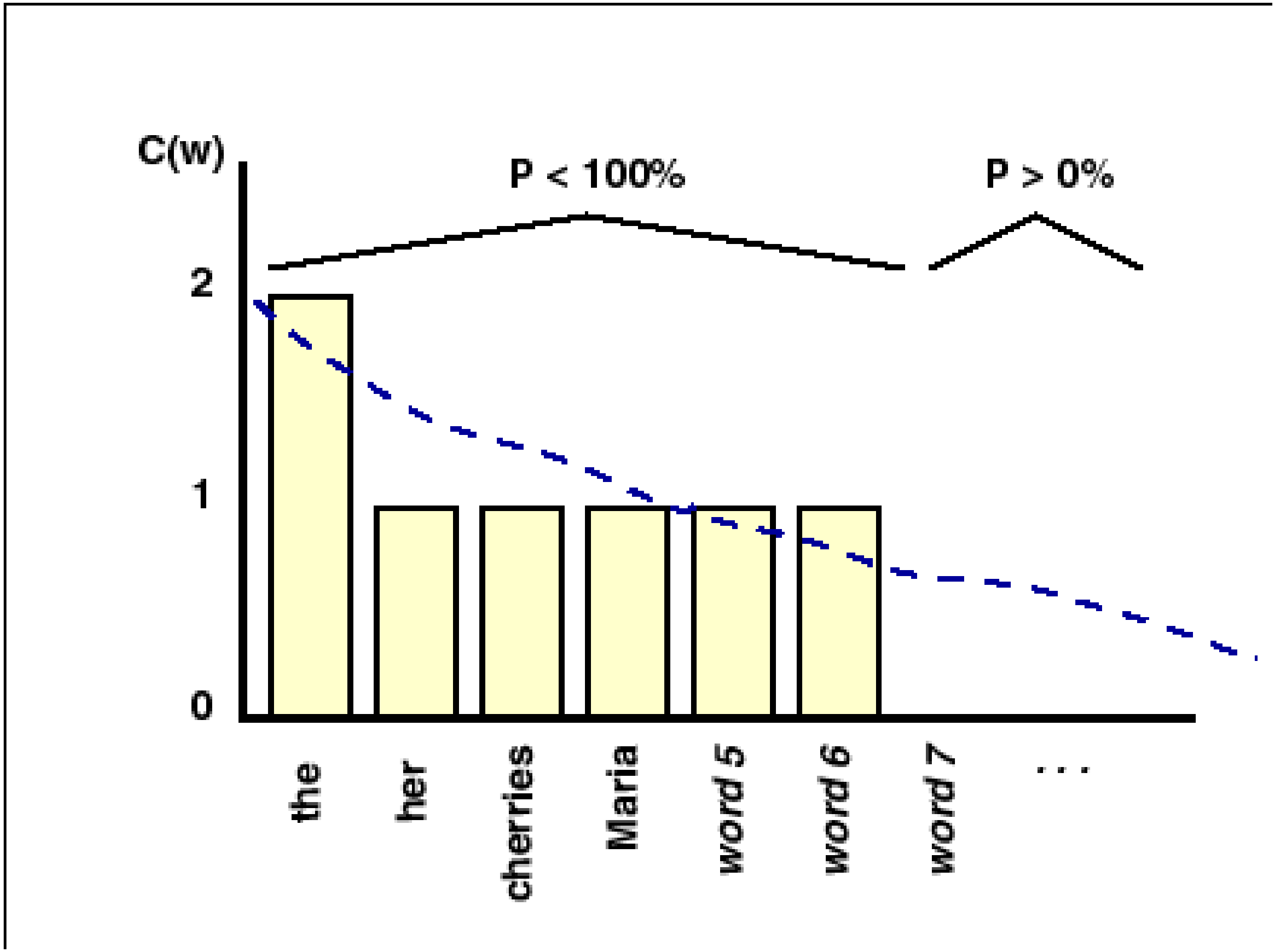
- Test set

  … denied the offer
  … denied the loan

Actual Probability Distribution:



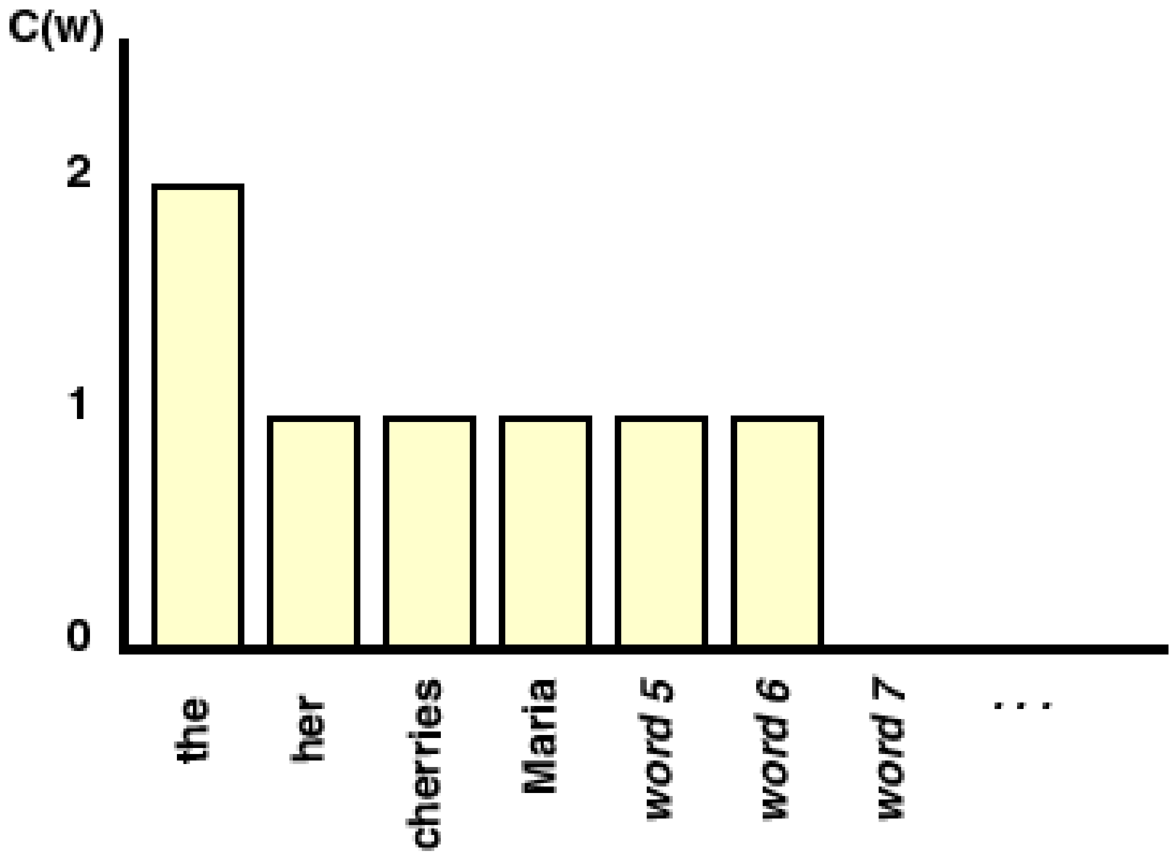
Actual Probability Distribution:



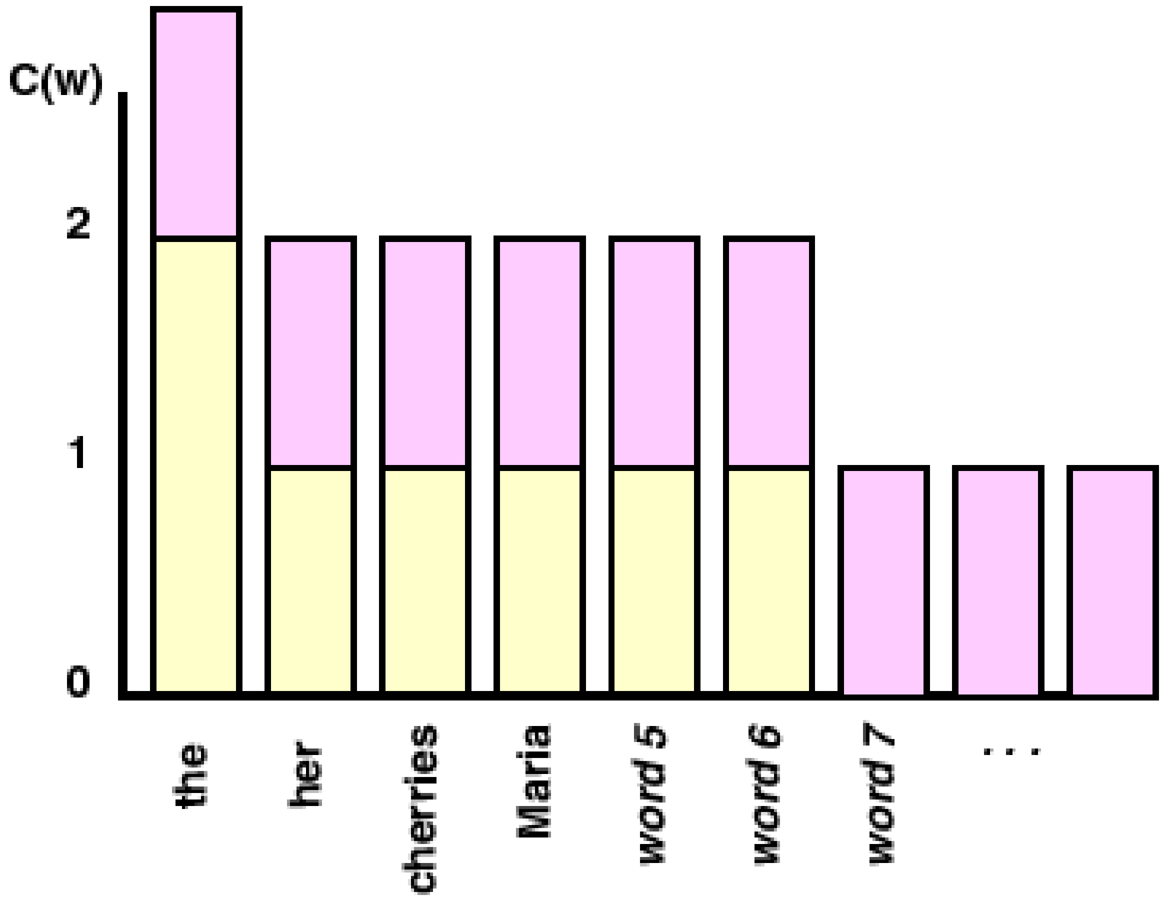
“Smoothing”

- Develop a model which decreases probability of seen events and allows the occurrence of previously unseen n-grams
- a.k.a. “Discounting methods”
- “Validation” – Smoothing methods which utilize a second batch of test data.

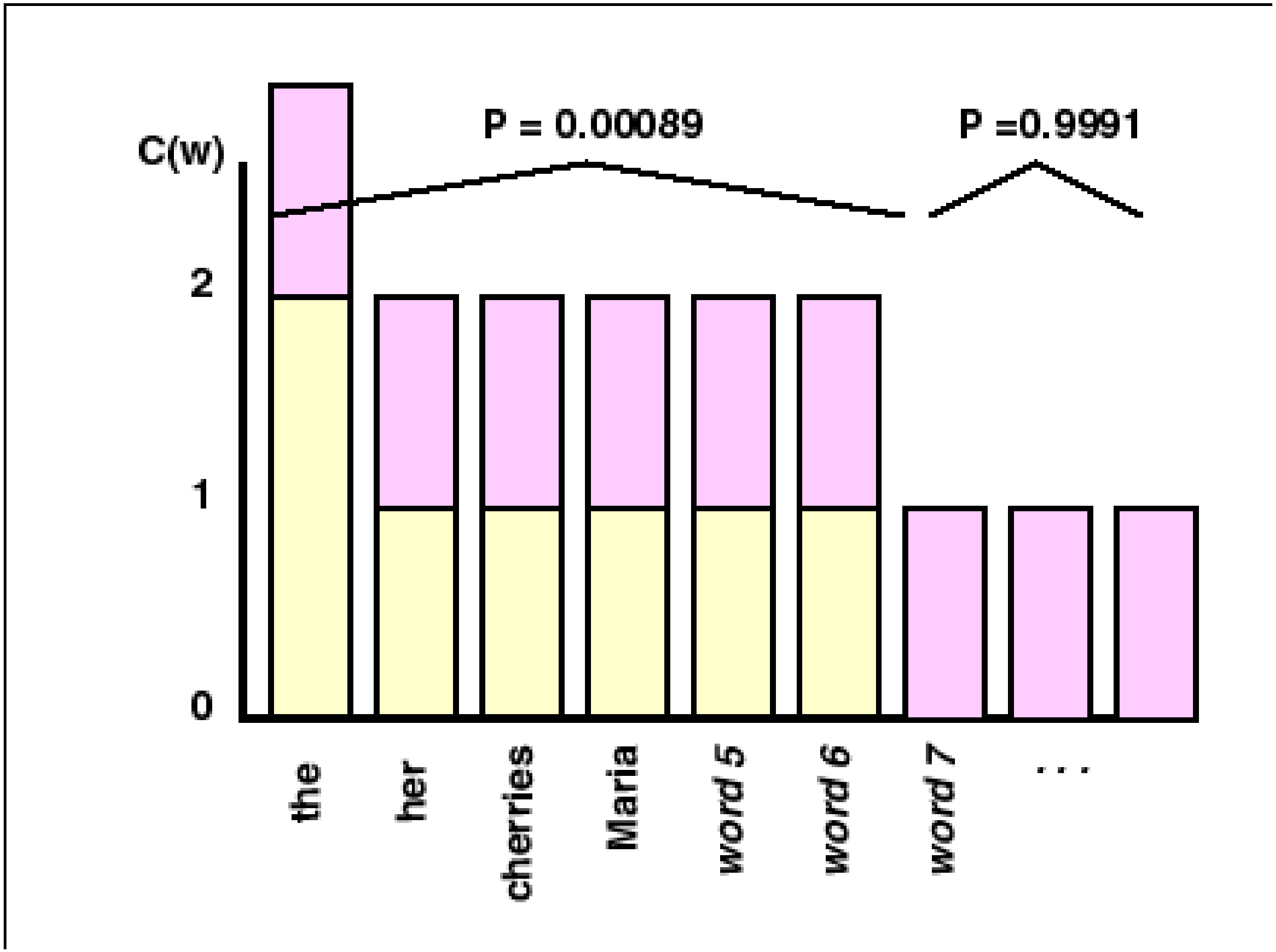
Smoothing



Smoothing: +1



Smoothing: +1


C(w)
P = 0.00089
P =0.9991
2
1
0
the
her
cherries
Maria
word 5
word 6
word 7
. . .

+◌ेन्द्र or = ◌ा + ◌िन्द्र ?

+◌ेन्द्र (1575):

- राजेन्द्र 137 ← **राज+** 978 ? □ **राजा+** 874;
  राजनीतिक 2236, राजनीति 1537, राज्य 5532
- नरेन्द्र 124 ← **नर+** 41,
  नरसिंह 40, नरक 37, नर्मदा 35, नर्सिंग 31, नरूला 30
- महेन्द्र 88
  ← **मह+** 0
  महिला 2682, महीने 2276, महसूस 856, महंगाई 737, महतो 645

  ← **महा+** 33 महाराष्ट्र 794, महासचिव 794, महान 400, महात्मा 275,
  महानिदेशक 199, महाराज 182, महानगर 179

  ?? महेश 283, महोत्सव 161

- note: केन्द्र 680 क 164, के 261214 की 163858 को 120489

LANGUAGE MODELING

Estimating N-gram Probabilities

Probabilistic Language Modeling

- Goal: determine if a sentence or phrase has a high acceptability in the language

  → compute the probability of the sequence of words

  E.g. "its water is so transparent that"

  - P(its, water, is, so, transparent, that)

Probabilistic Language Modeling

$$P(W) = P(w_1,w_2,w_3,w_4,w_5 \ldots w_n)$$

- Related task: probability of an upcoming word:

$$P(w_5 \mid w_1,w_2,w_3,w_4)$$

Reliability vs. Discrimination

- larger n: more information about the context of the specific instance (greater discrimination)

- smaller n: more instances in training data, better statistical estimates (more reliability)

How to compute P(W)

- Intuition: let’s rely on the Chain Rule of Probability

The Chain Rule

- Recall the definition of conditional probabilities:
  $P(B|A) = P(A,B) / P(A) \rightarrow$
  $$P(A,B) = P(A)\,P(B|A) \qquad [\text{Assume: } P(A) > 0]$$

- More variables:
  $$P(A,B,C,D) = P(A)\,P(B|A)\,P(C|A,B)\,P(D|A,B,C)$$
  Proof: Induction on the form:
  $$P((A,B),C)) = P(A,B)\,P(C|(A,B)) = P(A)\,P(B|A)\,P(C|A,B)$$

The Chain Rule

- Chain Rule in General

$$P(x_1,x_2,x_3,\ldots,x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1,\ldots,x_{n-1})$$

- Proof:
  - Holds for n=2 (Product rule)
  - Assume is true for $X = x_1 \ldots x_{n-1}$.

$$P(X, x_n) = P(X)\, P(x_n|X) \rightarrow \text{General chain rule}$$

The Chain Rule

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

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)

The Chain Rule

- Chain Rule in General

$$P(x_1,x_2,x_3,\ldots,x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1,\ldots,x_{n-1})$$

- Most useful when dependency of x_k is limited to only a few recent terms
  - First-order Markovian: x_k depends only on x_{k-1}

Estimating the probabilities

- Could we just count and divide?

$$P(\text{the} \mid \text{its water is so transparent that}) = \frac{Count(\text{its water is so transparent that the})}{Count(\text{its water is so transparent that})}$$

- Unlikely to find ANY instances in corpus!

Markov Assumption

Andrei Markov
1856-1922, Russia

- Simplifying assumption:

  Depends only on *k*-nearby text

- *First-order* Markov Process (k= 1):

$$P(\text{the} \mid \text{its water is so transparent that}) \text{ » } P(\text{the} \mid \text{that})$$

- or *Second-order* (k=2):

$$P(\text{the} \mid \text{its water is so transparent that}) \text{ » } P(\text{the} \mid \text{transparent that})$$

Markov Assumption

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

- In other words, we approximate each component in the product

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

Estimating bigram probabilities

- The Maximum Likelihood Estimate

$$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

Sentence Genration

Unigram Model: No dependencies on previous words

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i)$$

Bigram Model : Depends on 1 previous word

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

The Shannon Generation Method

- Choose a random bigram
  (<s>, w) according to its probability
- Now choose a random bigram
  (w, x) according to its probability
- And so on until we choose </s>
- Then string the words together

<s> I
I want
want to
to eat
eat Chinese
Chinese food
food </s>
I want to eat Chinese food

Shannon generation: Letters

- 1. Zero-order
  - XFOML RXKHR JFF JU J ZLPWCFWKCY JFFJEYVKCQSGXYD QI'AAMKBZAACIBZLHJQD
- 2. First-order (unigram frequencies as English)
  - OCR0 HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENH'ITPA OOBTTVA NAH BRL
- 3. Second-order (bigram).
  - ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE
- 4. Third-order (trigram)
  - IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

Shannon generation: Words

- 5. Word models: First-Order
  - REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE
- 6. Word Model: Second-Order (bigram)
  - THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED T

Claude E. Shannon. *A Mathematical Theory of Communication*, 1948.

The Corpus matters

- What corpus was used to generate these:

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.

The Corpus matters

- What corpus was used to generate these:

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

N-gram frequency falls rapidly w N

- Shakespeare Corpus: N=884,647 tokens, V=29,066
- Shakespeare produced 300,000 bigram types out of V^2= 844 million possible bigrams.
  - So 99.96% of the possible bigrams were never seen (have zero entries in the table)
- Quadrigrams worse: Shakespeare had very specific patterns of usage

Limitations of N-gram models

- Advantages:
  - Does not require expensive annotated corpora
  - Annotations are often disputed
  - Efficacy of intermediate representations are doubtful
- We can extend to trigrams, 4-grams, 5-grams
  - Corpus size must grow exponentially larger
- Main Disadvatage: **Long-distance dependencies**:

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

Practical Issues

- We do everything in log space
  - Avoid underflow
  - (also adding is faster than multiplying)

$$\log(p_1 \times p_2 \times p_3 \times p_4) = \log p_1 + \log p_2 + \log p_3 + \log p_4$$

Google N-Gram Release, August 2006

AUG
3

All Our N-gram are Belong to You

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects,

…

That's why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

Google N-Gram Release

- serve as the incoming 92
- serve as the incubator 99
- serve as the independent 794
- serve as the index 223
- serve as the indication 72
- serve as the indicator 120
- serve as the indicators 45
- serve as the indispensable 111
- serve as the indispensible 40
- serve as the individual 234

http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

Google N-Gram Release, August 2006

AUG

3

All Our N-gram are Belong to You

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects,

…

That's why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

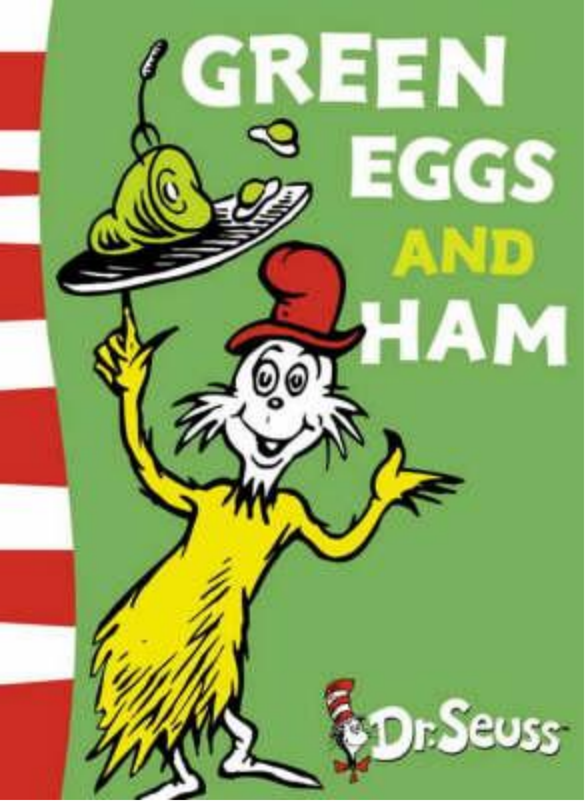
Google N-Gram Release

- serve as the incoming 92
- serve as the incubator 99
- serve as the independent 794
- serve as the index 223
- serve as the indication 72
- serve as the indicator 120
- serve as the indicators 45
- serve as the indispensable 111
- serve as the indispensible 40
- serve as the individual 234

http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

Google Book N-grams

- http://ngrams.googlelabs.com/



CS 671 NLP
LANGUAGE MODELING:
N-GRAMS

amitabha mukerjee
iit kanpur

Reliability vs. Discrimination

- larger n: more information about the context of the specific instance (greater discrimination)

- smaller n: more instances in training data, better statistical estimates (more reliability)

How to compute P(W)

- How to compute this joint probability:
  - P(its, water, is, so, transparent, that)
- Intuition: let’s rely on the Chain Rule of Probability