Lecture 6

Language Modeling/Pronunciation Modeling

Michael Picheny, Bhuvana Ramabhadran, Stanley F. Chen, Markus Nussbaum-Thom

Watson Group
IBM T.J. Watson Research Center
Yorktown Heights, New York, USA
{picheny,bhuvana,stanchen,nussbaum}@us.ibm.com

24 February 2016

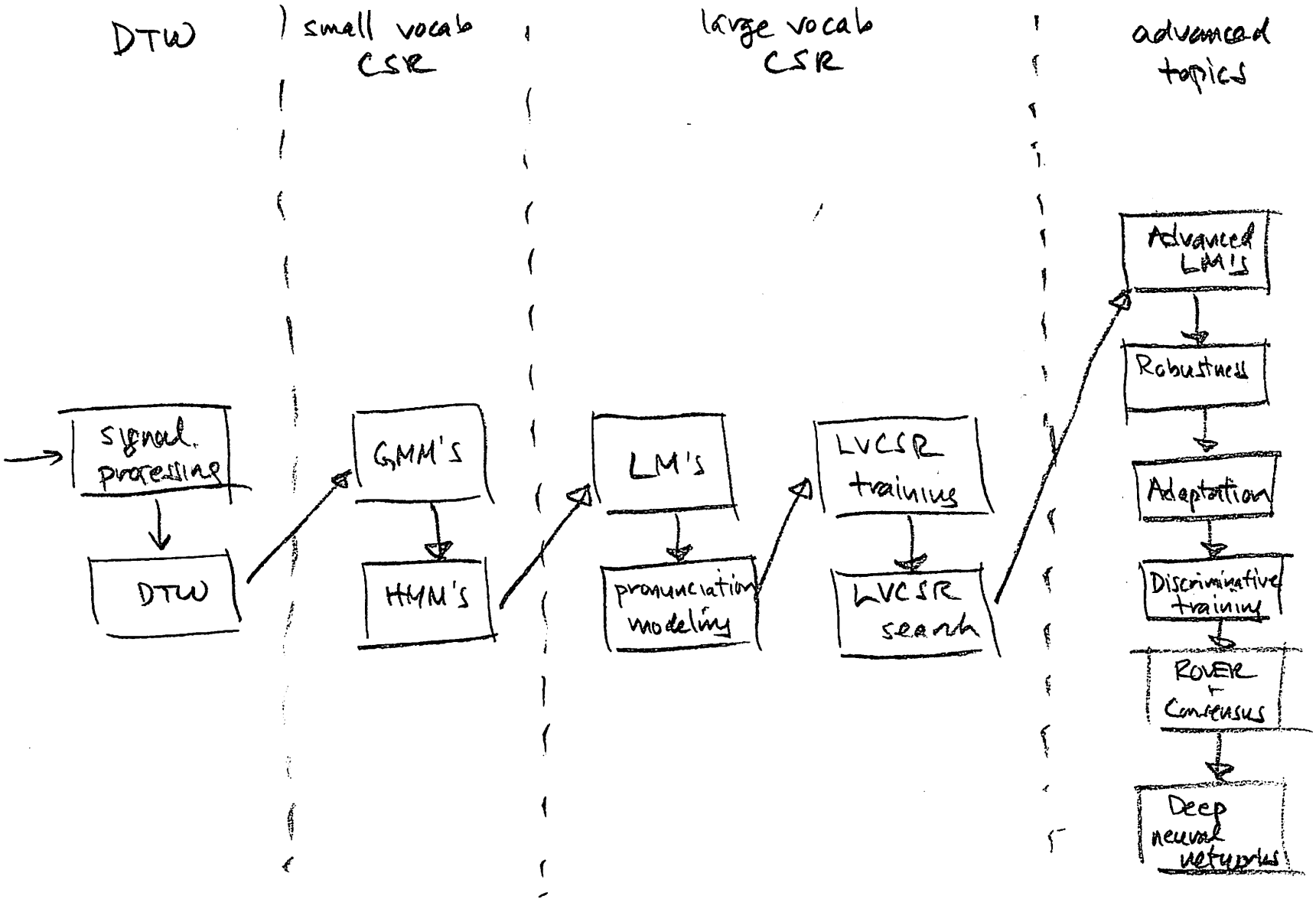
Administrivia

- Complete lecture 4+ slides posted.
- Lab 1
  - Handed back today?
  - Sample answers:
    `/user1/faculty/stanchen/e6870/lab1_ans/`
  - Awards ceremony.
- Lab 2
  - Due two days from now (Friday, Feb. 26) at 6pm.
  - Piazza is your friend.
  - Remember: two free extension days for *one* lab.
- Lab 3 posted by Friday.

Feedback

- Clear (8), mostly clear (5).
- Pace: fast (2), OK (3).
- Muddiest: HMM's (2), decoding (1), continuous ASR (1), silence (1), posterior counts (1), ϵ arcs (1), training (1).
- Comments (2+ votes)
  - Demos good (5)
  - Need more time for lab 2 (4)
  - Lots of big picture info, connecting everything good (2)
  - Good diagrams (2)

Road Map


DTW
small vocab CSR
large vocab CSR
advanced topics
signal processing
DTW
GMM's
HMM's
LM's
pronunciation modeling
LVCSR training
LVCSR search
Advanced LM's
Robustness
Adaptation
Discriminative training
ROVER + Consensus
Deep neural networks

Review, Part I

- What is $\mathbf{x}$?
  - The feature vector.
- What is ω?
  - A word sequence.
- What notation do we use for acoustic models?
  - $P(\mathbf{x}|\omega)$
- What does an acoustic model model?
  - How likely feature vectors are given a word sequence.
- What notation do we use for language models?
  - $P(\omega)$
- What does a language model model?
  - How frequent each word sequence is.

Review, Part II

- How do we do DTW recognition?

$$\text{(answer)} = ???$$

$$\text{(answer)} = \arg\max_{\omega \in \text{vocab}} P(\mathbf{x}|\omega)$$

- What is the fundamental equation of ASR?

$$\begin{aligned}
\text{(answer)} &= \arg\max_{\omega \in \text{vocab}^*} \text{(language model)} \times \text{(acoustic model)} \\
&= \arg\max_{\omega \in \text{vocab}^*} \text{(prior prob over words)} \times P(\text{feats}|\text{words}) \\
&= \arg\max_{\omega \in \text{vocab}^*} P(\omega) P(\mathbf{x}|\omega)
\end{aligned}$$

How Do Language Models Help?

$$\begin{aligned}(\text{answer}) &= \arg\max_{\omega} \ (\text{language model}) \times (\text{acoustic model}) \\ &= \arg\max_{\omega} \ P(\omega)P(\mathbf{x}|\omega)\end{aligned}$$

- Homophones.

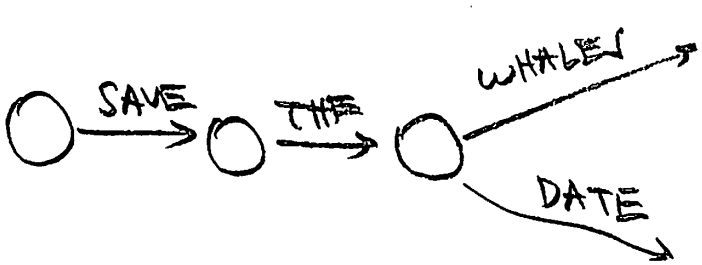
  THIS IS OUR ROOM FOR A FOUR HOUR PERIOD .
  THIS IS HOUR ROOM FOUR A FOR OUR . PERIOD

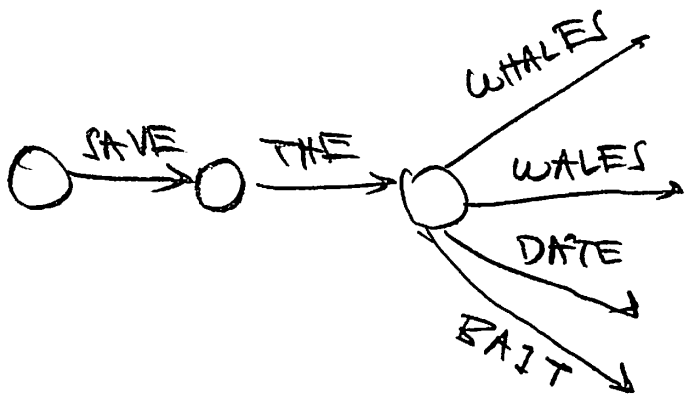
- Confusable sequences in general.

  IT IS EASY TO RECOGNIZE SPEECH .
  IT IS EASY TO WRECK A NICE PEACH .

Language Modeling: Goals

- Assign high probabilities to the good stuff.
- Assign low probabilities to the bad stuff.
  - Restrict choices given to AM.





Part I

Language Modeling

Where Are We?

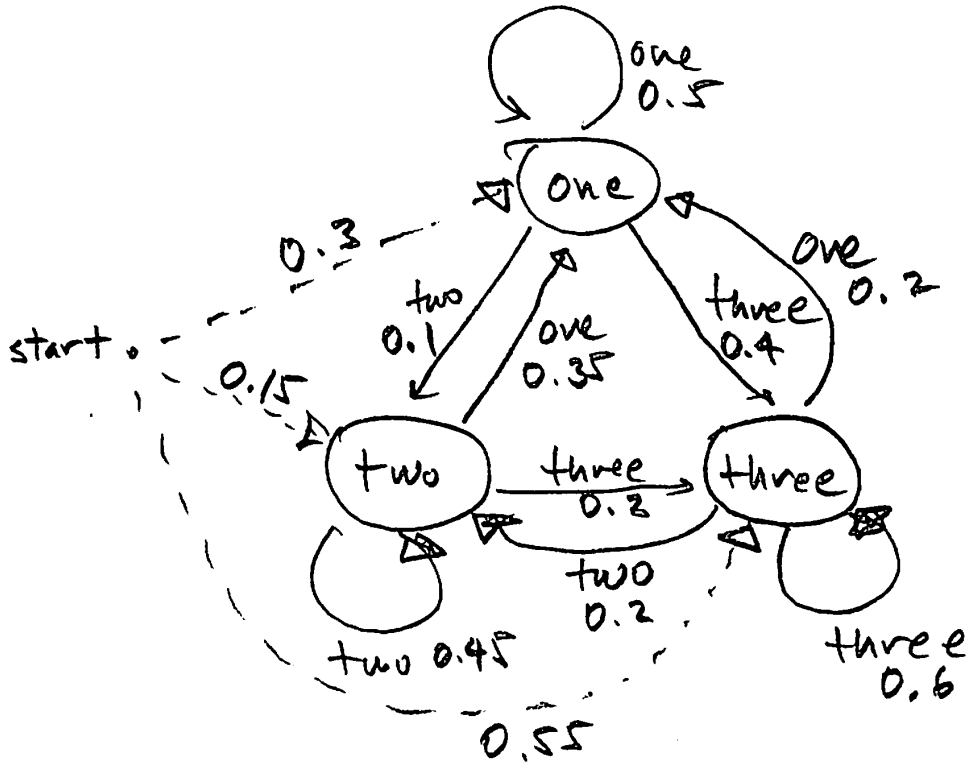
Let's Design a Language Model!

- Goal: probability distribution over word sequences.

$$P(\omega) = P(w_1 w_2 \cdots)$$

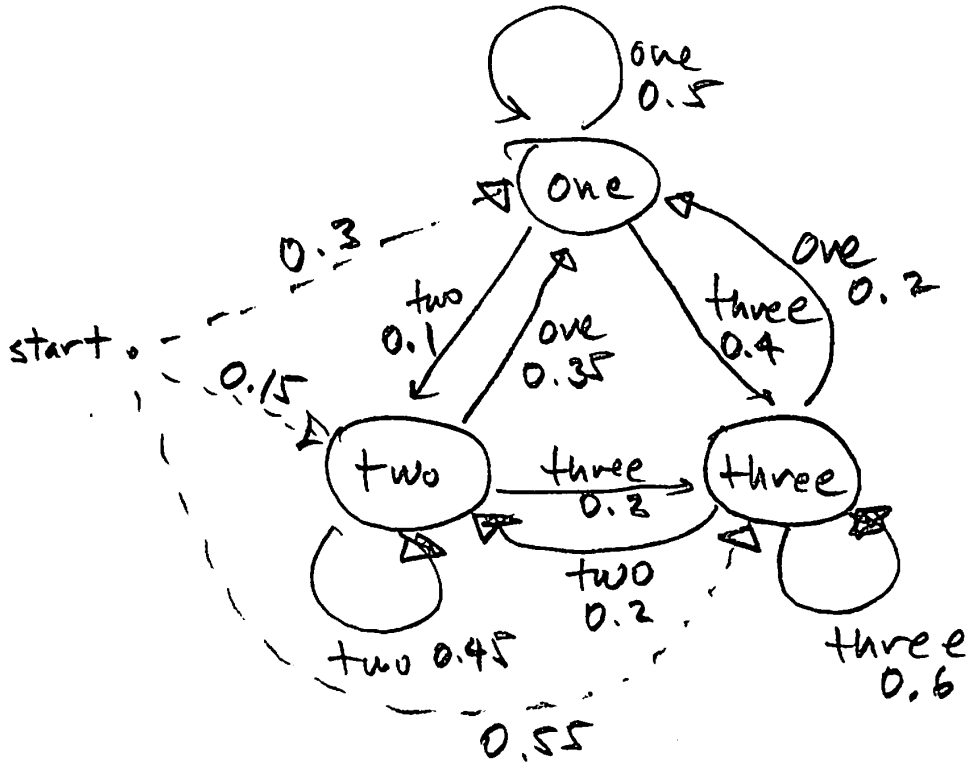
- What type of model? What's the simplest we can do?

Markov Model, Order 1: Bigram Model



- State $\Leftrightarrow$ last word.
- Sum of arc probs leaving state is 1.

Bigram Model Example



$P(\text{one three two}) = ???$

$$P(\text{one three two}) = 0.3 \times 0.4 \times 0.2 = 0.024$$

Bigram Model Equations

$$P(\text{one three two}) = P(\text{one}) \times P(\text{three}|\text{one}) \times P(\text{two}|\text{three})$$
$$= 0.3 \times 0.4 \times 0.2 = 0.024$$

$$P(w_1, \ldots, w_L) = \prod_{i=1}^{L} P(\text{cur word} \mid \text{last word})$$
$$= \prod_{i=1}^{L} P(w_i|w_{i-1})$$

What Training Data?

- Text! As a list of *utterances*.

  I WANT TO FLY FROM AUSTIN TO BOSTON
  CAN I GET A VEGETARIAN MEAL
  DO YOU HAVE ANYTHING THAT IS NONSTOP
  I WANT TO LEAVE ON FEBRUARY TWENTY SEVEN
  WHO LET THE DOGS OUT
  GIVE ME A ONE WAY TICKET PAUSE TO HELL

- Are AM's or LM's usually trained with more data?

Incomplete Utterances

- Example: I'M GOING TO

$$P(\text{I'M}) \times P(\text{GOING}|\text{I'M}) \times P(\text{TO}|\text{GOING})$$

- Is this a good utterance?
- Does this get a good score?
- How to fix this?

Utterance Begins and Ends

- Add beginning-of-sentence token; *i.e.*, $w_0 = \triangleright$.
- Predict end-of-sentence token at end; *i.e.*, $w_{L+1} = \triangleleft$.

$$P(w_1 \cdots w_L) = \prod_{i=1}^{L+1} P(w_i | w_{i-1})$$

- Does this fix problem?

$$P(\text{I'M GOING TO}) = P(\text{I'M}|\triangleright) \times P(\text{GOING}|\text{I'M}) \times P(\text{TO}|\text{GOING}) \times P(\triangleleft|\text{TO})$$

*Side effect: $\sum_\omega P(\omega) = 1$. (Can you prove this?)

How to Set Probabilities?

- How to estimate $P(\text{FLY}|\text{TO})$?

$$P(\text{FLY}|\text{TO}) = \frac{\text{count}(\text{TO FLY})}{\text{count}(\text{TO})}$$

- MLE: count and normalize!

$$P_{\text{MLE}}(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_w c(w_{i-1}w)} = \frac{c(w_{i-1}w_i)}{c(w_{i-1})}$$

Example: Maximum Likelihood Estimation

- 23M words of Wall Street Journal text.

  FEDERAL HOME LOAN MORTGAGE CORPORATION –DASH ONE .POINT FIVE BILLION DOLLARS OF REALESTATE MORTGAGE -HYPHEN INVESTMENT CONDUIT SECURITIES OFFERED BY MERRILL LYNCH &AMPERSAND COMPANY .PERIOD

  NONCOMPETITIVE TENDERS MUST BE RECEIVED BY NOON EASTERN TIME THURSDAY AT THE TREASURY OR AT FEDERAL RESERVE BANKS OR BRANCHES .PERIOD

  THE PROGRAM ,COMMA USING SONG ,COMMA PUPPETS AND VIDEO ,COMMA WAS CREATED AT LAWRENCE LIVERMORE NATIONAL LABORATORY ,COMMA LIVERMORE ,COMMA CALIFORNIA ,COMMA AFTER A PARENT AT A BERKELEY ELEMENTARY SCHOOL EXPRESSED INTEREST .PERIOD

Example: Bigram Model

- $P(\text{I HATE TO WAIT}) = ???$
- $P(\text{EYE HATE TWO WEIGHT}) = ???$
- Step 1: Collect all bigram counts, unigram history counts.

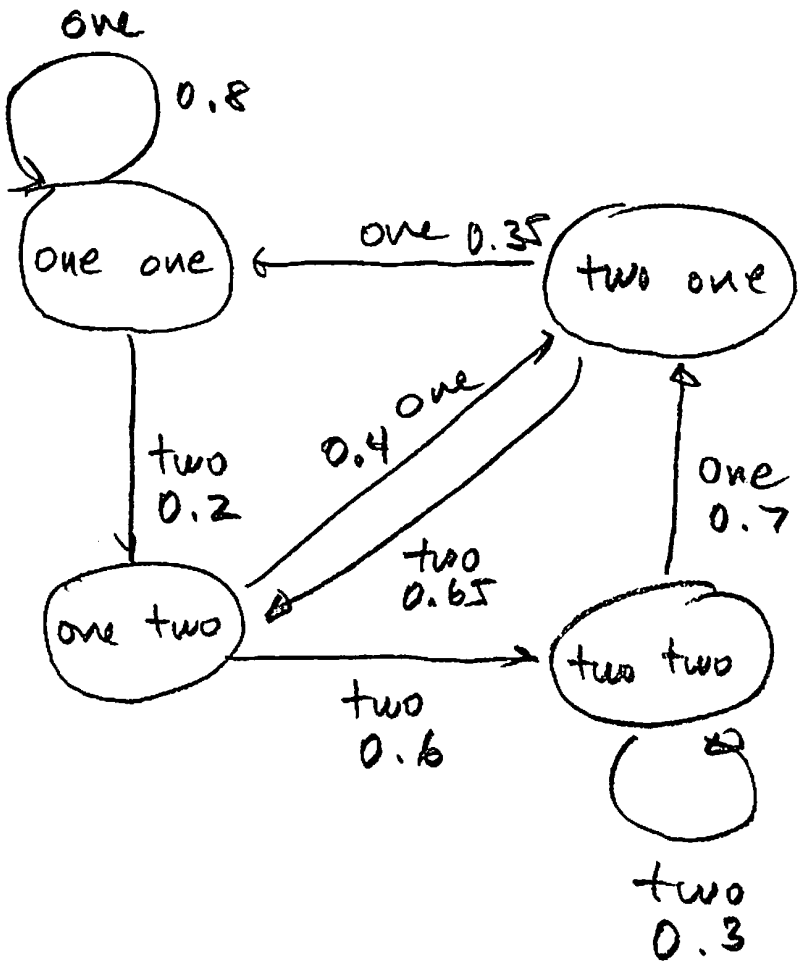
	EYE	I	HATE	TO	TWO	WAIT	WEIGHT	$\triangleleft$	$*$
$\triangleright$	3	3234	5	4064	1339	8	22	0	892669
EYE	0	0	0	26	1	0	0	52	735
I	0	0	45	2	1	1	0	8	21891
HATE	0	0	0	40	0	0	0	9	246
TO	8	6	19	21	5341	324	4	221	510508
TWO	0	5	0	1617	652	0	0	4213	132914
WAIT	0	0	0	71	2	0	0	35	882
WEIGHT	0	0	0	38	0	0	0	45	643

Example: Bigram Model

$$
\begin{aligned}
&P(\text{I HATE TO WAIT}) \\
&= P(\text{I}|\triangleright)P(\text{HATE}|\text{I})P(\text{TO}|\text{HATE})P(\text{WAIT}|\text{TO})P(\triangleleft|\text{WAIT}) \\
&= \frac{3234}{892669} \times \frac{45}{21891} \times \frac{40}{246} \times \frac{324}{510508} \times \frac{35}{882} = 3.05 \times 10^{-11}
\end{aligned}
$$

$$
\begin{aligned}
&P(\text{EYE HATE TWO WEIGHT}) \\
&= P(\text{EYE}|\triangleright)P(\text{HATE}|\text{EYE})P(\text{TWO}|\text{HATE})P(\text{WEIGHT}|\text{TWO}) \times \\
&\qquad P(\triangleleft|\text{WEIGHT}) \\
&= \frac{3}{892669} \times \frac{0}{735} \times \frac{0}{246} \times \frac{0}{132914} \times \frac{45}{643} = 0
\end{aligned}
$$

What's Better Than First Order?



$P(\text{two} \mid \text{one two}) = ???$

Bigram *vs.* Trigram

- Bigram

$$P(\text{SAM I AM}) = P(\text{SAM}|\triangleright)P(\text{I}|\text{SAM})P(\text{AM}|\text{I})P(\triangleleft|\text{AM})$$

$$P(\text{AM}|\text{I}) = \frac{c(\text{I AM})}{c(\text{I})}$$

- Trigram

$$P(\text{SAM I AM}) = P(\text{SAM}|\triangleright\triangleright)P(\text{I}|\triangleright\,\text{SAM})P(\text{AM}|\text{SAM I})P(\triangleleft|\text{I AM})$$

$$P(\text{AM}|\text{SAM I}) = \frac{c(\text{SAM I AM})}{c(\text{SAM I})}$$

Markov Model, Order 2: Trigram Model

$$P(w_1, \ldots, w_L) = \prod_{i=1}^{L+1} P(\text{cur word} \mid \text{last 2 words})$$
$$= \prod_{i=1}^{L+1} P(w_i | w_{i-2} w_{i-1})$$

$$P(w_i | w_{i-2} w_{i-1}) = \frac{c(w_{i-2} w_{i-1} w_i)}{c(w_{i-2} w_{i-1})}$$

Recap: *N*-Gram Models

- Markov model of order $n-1$.
  - Predict current word from last $n-1$ words.
- Don't forget utterance begins and ends.
- Easy to train: count and normalize.
- Easy as pie.

Pop Quiz

- How many states are in the HMM for a unigram model?
- What do you call a Markov Model of order 3?

Where Are We?

Zero Counts

- THE WORLD WILL END IN TWO THOUSAND THIRTY EIGHT
- What if $c(\text{TWO THOUSAND THIRTY}) = 0$?

$$P(w_1, \ldots, w_L) = \prod_{i=1}^{L+1} P(\text{cur word} \mid \text{last 2 words})$$

$$(\text{answer}) = \arg\max_{\omega} \ (\text{language model}) \times (\text{acoustic model})$$

- Goal: assign high probabilities to the good stuff!?

the hour by **bigbirdz**. Some rights reserved.

How Bad Is the Zero Count Problem?

- Training set: 11.8M words of WSJ.
- In held-out WSJ data, what fraction trigrams unseen?
  - $<$5%
  - 5–10%
  - 10–20%
  - $>$20%
- 36.6%!

Zero Counts, Visualized

BUT THERE'S MORE .PERIOD

IT'S NOT LIMITED TO PROCTER .PERIOD

MR. ANDERS WRITES ON HEALTH CARE FOR THE JOURNAL .PERIOD

ALTHOUGH PEOPLE'S PURCHASING POWER HAS FALLEN AND SOME HEAVIER INDUSTRIES ARE SUFFERING ,COMMA FOOD SALES ARE GROWING .PERIOD

"DOUBLE-QUOTE THE FIGURES BASICALLY SHOW THAT MANAGERS HAVE BECOME MORE NEGATIVE TOWARD U. S. EQUITIES SINCE THE FIRST QUARTER ,COMMA "DOUBLE-QUOTE SAID ANDREW MILLIGAN ,COMMA AN ECONOMIST AT SMITH NEW COURT LIMITED .PERIOD

P. &AMPERSAND G. LIFTS PRICES AS OFTEN AS WEEKLY TO COMPENSATE FOR THE DECLINE OF THE RUBLE ,COMMA WHICH HAS FALLEN IN VALUE FROM THIRTY FIVE RUBLES TO THE DOLLAR IN SUMMER NINETEEN NINETY ONE TO THE CURRENT RATE OF SEVEN HUNDRED SIXTY SIX .PERIOD

Maximum Likelihood and Sparse Data

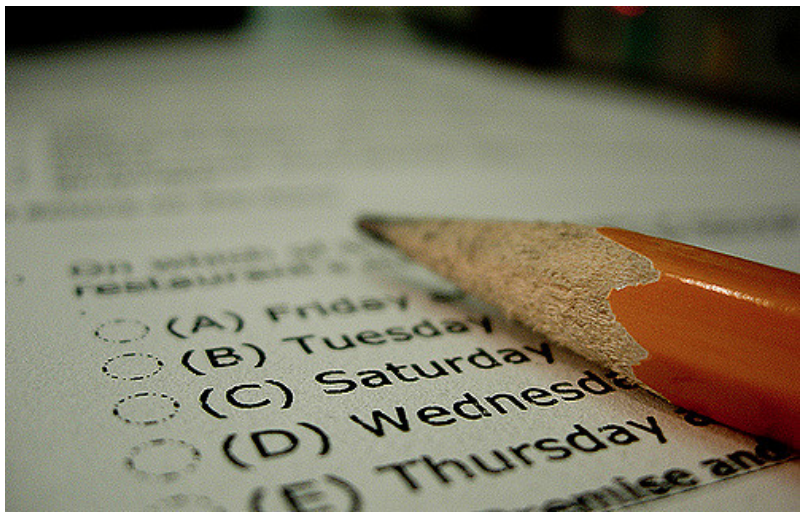
- In theory, ML estimate is as good as it gets . . .
  - In limit of lots of data.
- In practice, sucks when data is *sparse*.
  - Bad for n-grams with zero or low counts.
- All n-gram models are sparse!

MLE and 1-counts

- Training set: 11.8M word of WSJ.
- Test set: 11.8M word of WSJ.
- If trigram has 1 count in training set . . .
- How many counts does it have on average in test?
  - >0.9
  - 0.5–0.9
  - 0.3–0.5
  - <0.3
- 0.22! *i.e.*, MLE is off by factor of 5!

Smoothing

- MLE $\Leftrightarrow$ frequency of *n*-gram in training data!
- Goal: estimate frequencies of *n*-grams in test data!
- Smoothing $\Leftrightarrow$ *regularization*.
  - Adjust ML estimates to better match test data.



Where Are We?

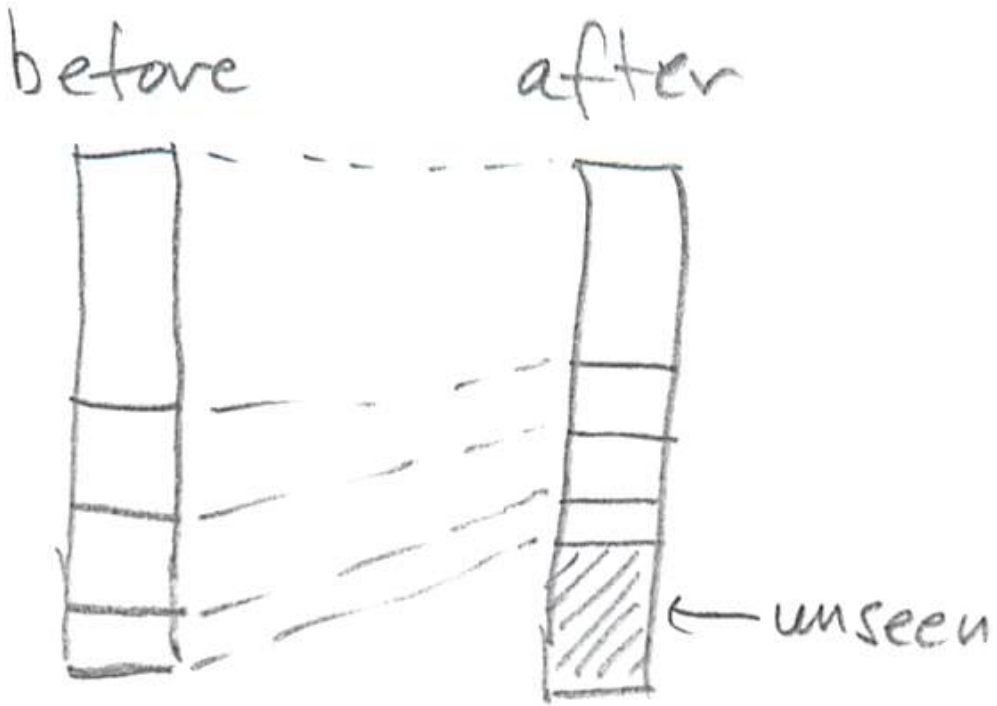
Baseline: MLE Unigram Model

$$P_{\text{MLE}}(w) = \frac{c(w)}{\sum_{w} c(w)}$$

word	count	P_{MLE}
ONE	5	0.5
TWO	2	0.2
FOUR	2	0.2
SIX	1	0.1
ZERO	0	0.0
THREE	0	0.0
FIVE	0	0.0
SEVEN	0	0.0
EIGHT	0	0.0
NINE	0	0.0
total	10	1.0

The Basic Idea

- How to adjust probs of words with zero count, on average?
- How to adjust probs of words with nonzero count?



Smoothing 1.0: Avoiding Zero Probs

- $+\delta$ smoothing, *e.g.*, $+1$ smoothing.

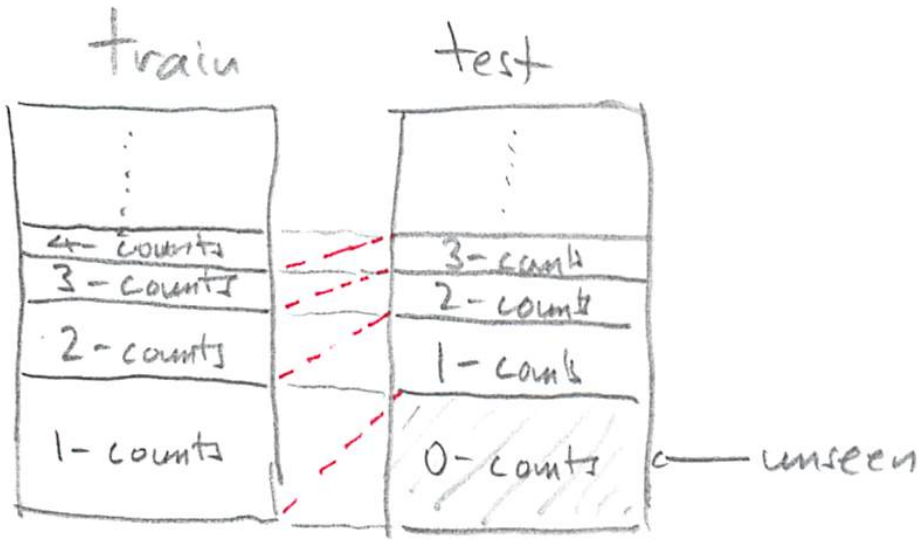
word	count	P_{MLE}	c_{smooth}	P_{smooth}
ONE	5	0.5	6	0.30
TWO	2	0.2	3	0.15
FOUR	2	0.2	3	0.15
SIX	1	0.1	2	0.10
ZERO	0	0.0	1	0.05
THREE	0	0.0	1	0.05
FIVE	0	0.0	1	0.05
SEVEN	0	0.0	1	0.05
EIGHT	0	0.0	1	0.05
NINE	0	0.0	1	0.05
total	10	1.0	20	1.0

Is This A Good Idea?

- Very sensitive to δ; how to set?
- Does it discount low and high counts correctly?
- Is there some principled way to do this?

The Good-Turing Estimate (1953)

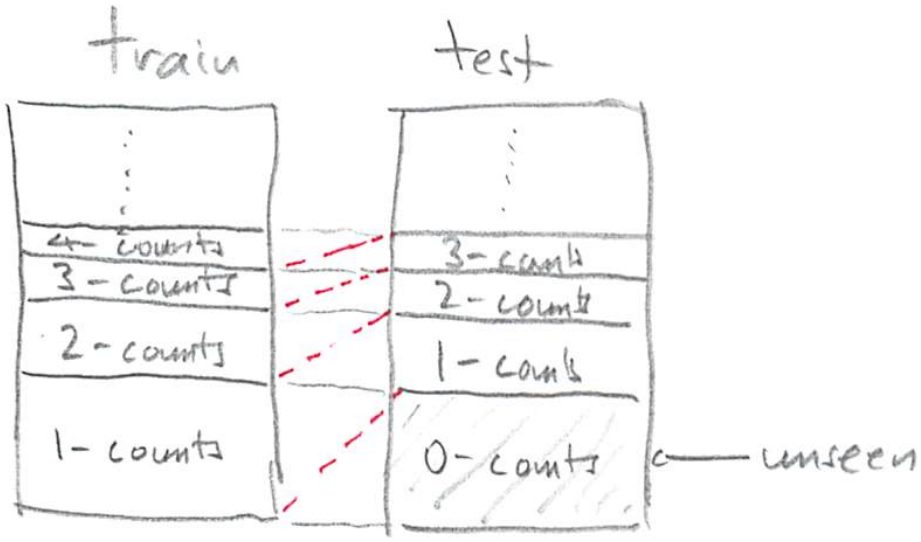
- Train set, test set of equal size.
- Total count mass of $(k+1)$-count words in training $\approx \ldots$
- Total count mass of k-count words in test!



What Is The Frequency of Unseen Trigrams?

- Total count mass of 1-count 3g's in training $\approx \ldots$
- Total count mass of 0-count (unseen) 3g's in test!
- Train: 4.32M 1-count 3g's $\Rightarrow$ 4.32M unseen test counts.
- Train/test set: 11.8M words.

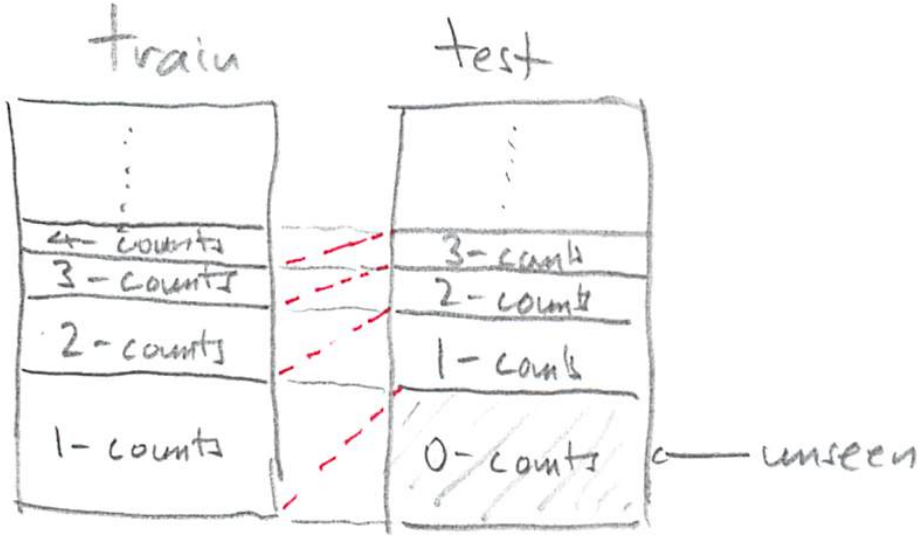
$$\frac{4.32\text{M}}{11.8\text{M}} = 36.6\%$$



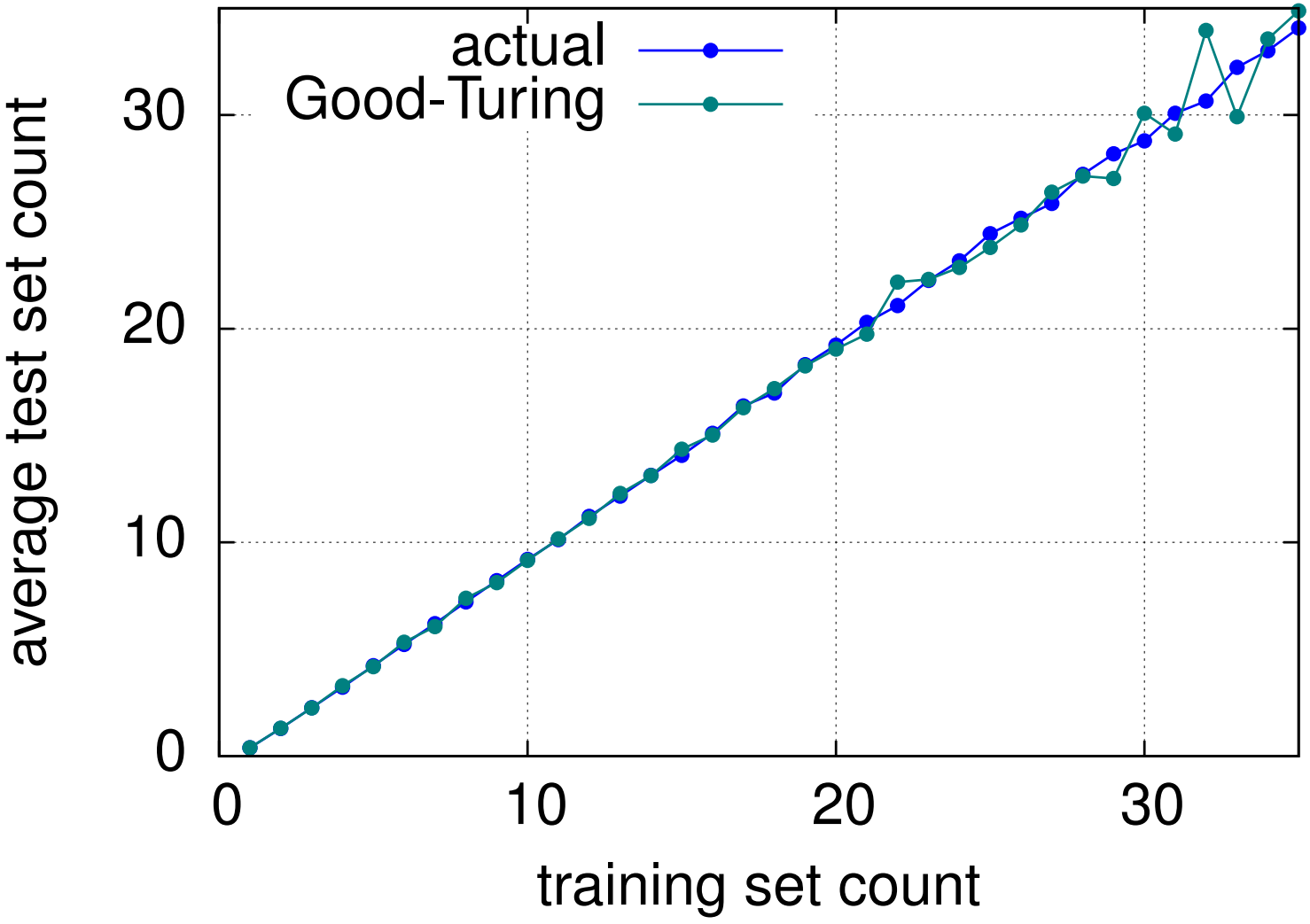
Smoothing Counts

$$P_{\text{MLE}}(w) = \frac{c(w)}{\sum_w c(w)} \quad \Rightarrow \quad P_{\text{smooth}}(w) = \frac{c_{\text{smooth}}(w)}{\sum_w c(w)}$$

$$c_{\text{smooth}}(k\text{-count word}) \approx \frac{(\#\text{ words w/ } k+1 \text{ counts}) \times (k+1)}{(\#\text{ words w/ } k \text{ counts})}$$



How Accurate Is Good-Turing?



Bigram counts; 11.8M words WSJ training and test.

Recap: Good-Turing

- Awesome way to smooth unigram models.
- Only works for counts with lots of counts.

$$c_{\text{smooth}}(k\text{-count word}) \approx \frac{(\#\text{ words w/ } k+1 \text{ counts}) \times (k+1)}{(\#\text{ words w/ } k \text{ counts})}$$

What About Bigram Models?

- Unigram $P(w_i) \Rightarrow$ bigram $P(w_i|w_{i-1})$.
- How to apply Good-Turing to bigrams?
- Idea 1:
  - Pool all bigrams; compute GT discounts.
  - Use smoothed counts to estimate probs.
- Idea 2:
  - Apply GT independently to each bigram distribution.
- Which idea makes more sense? Can we do better?

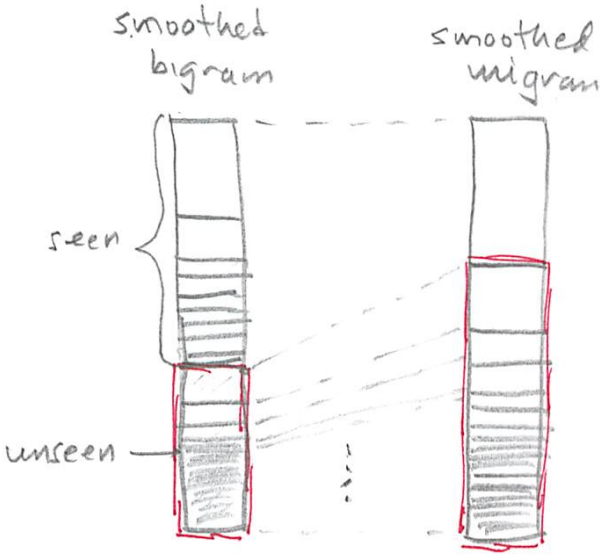
Case Study: Zero Counts

- Consider single 2g distribution: $P(w_i|\text{ANTONIO})$.
- $c(\text{ANTONIO THE}) = 0$, $c(\text{ANTONIO THEODORE}) = 0$.
- Does $P_{\text{GT}}(\text{THE}|\text{ANTONIO}) = P_{\text{GT}}(\text{THEODORE}|\text{ANTONIO})$?
- Can we do better?

Backoff

- Instead of assigning probs to unseen 2g's *uniformly* . . .
  - Assign proportionally to unigram distr $P(w_i)$!
  - *i.e.*, give more mass to THE than THEODORE.

$$P_{\text{smooth}}(w_i|w_{i-1}) = \begin{cases} P_{\text{GT}}(w_i|w_{i-1}) & \text{if } c(w_{i-1}w_i) > 0 \\ \alpha_{w_{i-1}} P_{\text{smooth}}(w_i) & \text{otherwise} \end{cases}$$



Putting It Together: Katz Smoothing (1987)

- If count high, no discounting (GT estimate unreliable).
- If count low, use GT estimate.
- If no count, use scaled backoff probability.

$$P_{\text{Katz}}(w_i|w_{i-1}) = \begin{cases} P_{\text{MLE}}(w_i|w_{i-1}) & \text{if } c(w_{i-1}w_i) \geq k \\ P_{\text{GT}}(w_i|w_{i-1}) & \text{if } 0 < c(w_{i-1}w_i) < k \\ \alpha_{w_{i-1}} P_{\text{Katz}}(w_i) & \text{otherwise} \end{cases}$$

Example: Katz Smoothing

- Conditional distribution: $P(w|\text{HATE})$.

w	c	P_{MLE}	c_{smooth}	P_{smooth}
TO	40	0.163	40.0000	0.162596
THE	22	0.089	20.9840	0.085301
IT	15	0.061	14.2573	0.057957
CRIMES	13	0.053	12.2754	0.049900
...	...	...	...	...
AFTER	1	0.004	0.4644	0.001888
ALL	1	0.004	0.4644	0.001888
...	...	...	...	...
A	0	0.000	1.1725	0.004766
AARON	0	0.000	0.0002	0.000001
...	...	...	...	...
total	246	1.000	246	1.000000

Recap: Smoothing

- ML estimates: way off for low counts.
  - Zero probabilities kill performance.
- Key aspects of smoothing algorithms.
  - How to discount counts of seen words.
  - Estimating mass of unseen words.
  - Backoff to get information from lower-order models.
- Katz smoothing was standard thru late 90's.

Did We Meet Our Language Model Goals?

- Assign high probabilities to the good stuff?
  - Seen *n*-grams get OK probs.
  - Unseen *n*-grams get not so bad probs.
- Assign low probabilities to the bad stuff?

Trigram Model, 20M Words of WSJ

AND WITH WHOM IT MATTERS AND IN THE SHORT -HYPHEN TERM

AT THE UNIVERSITY OF MICHIGAN IN A GENERALLY QUIET SESSION

THE STUDIO EXECUTIVES LAW

REVIEW WILL FOCUS ON INTERNATIONAL UNION OF THE STOCK MARKET

HOW FEDERAL LEGISLATION

"DOUBLE-QUOTE SPENDING

THE LOS ANGELES

THE TRADE PUBLICATION

SOME FORTY %PERCENT OF CASES ALLEGING GREEN PREPARING FORMS

NORTH AMERICAN FREE TRADE AGREEMENT (LEFT-PAREN NAFTA)RIGHT-PAREN ,COMMA WOULD MAKE STOCKS

A MORGAN STANLEY CAPITAL INTERNATIONAL PERSPECTIVE ,COMMA GENEVA

"DOUBLE-QUOTE THEY WILL STANDARD ENFORCEMENT

THE NEW YORK MISSILE FILINGS OF BUYERS

Where Are We?

This Section

- How do we evaluate ASR systems?
- Can we evaluate LM's outside of ASR?
- Can we evaluate AM's outside of ASR?

What is This Word Error Rate Thing?

- Most popular evaluation measure for ASR systems
- Divide number of errors by number of words.

$$\text{WER} \equiv \frac{\sum_{\text{utts } u}(\#\text{ errors in } u)}{\sum_{\text{utts } u}(\#\text{ words in reference for } u)}$$

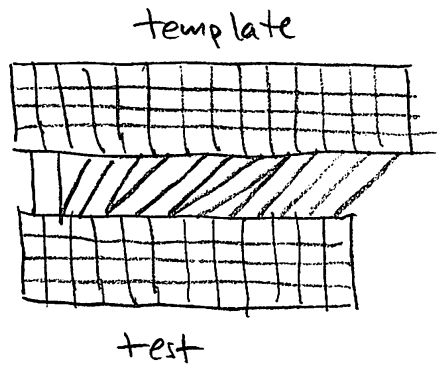
- What is “number of errors” in utterance?
- Minimum # insertions, deletions, and substitutions.

Example: Word Error Rate

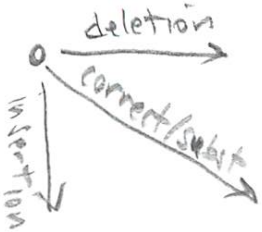
- What is the WER?

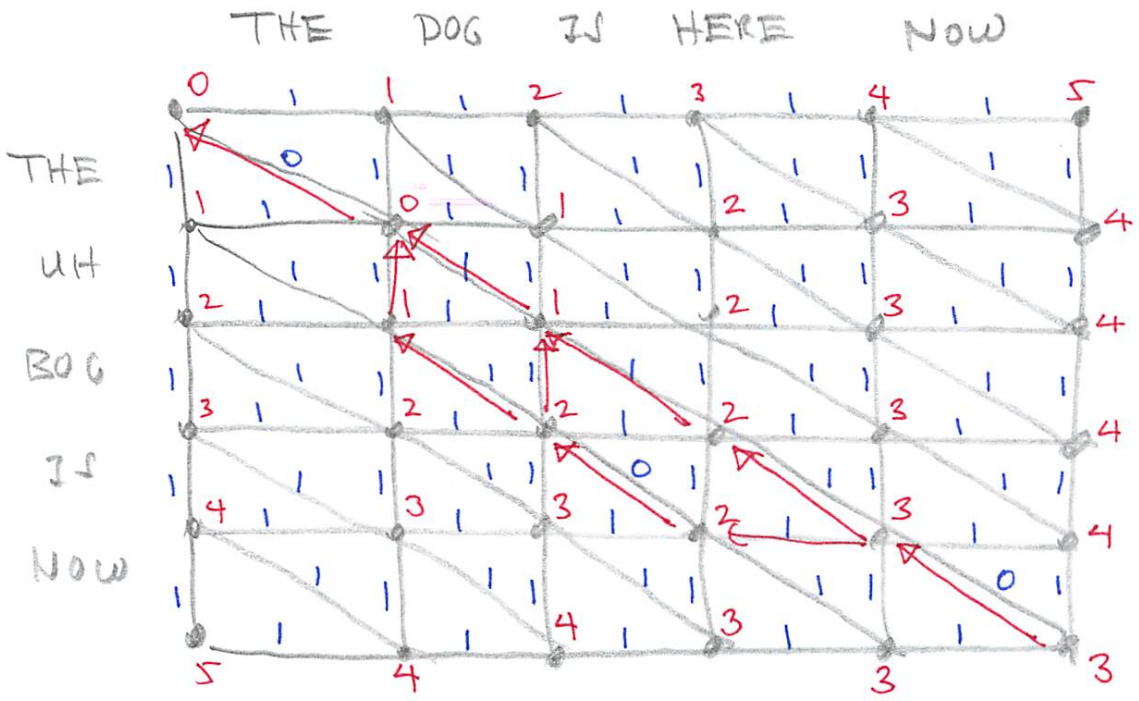
reference:	THE DOG IS HERE NOW
hypothesis:	THE UH BOG IS NOW

- Can WER be above 100%?
- What algorithm to compute WER?
- Dynamic programming/DTW.



Computing Word Error Rate





Getting a Feel For Word Error Rates

- 0% WER

  SAYS DAVIS DYER ,COMMA A WINTHROP GROUP MANAGING DIRECTOR :COLON "DOUBLE-QUOTE MY FIRST STOP IS TO GO TO THE FACTORY AND SAY ,COMMA 'SINGLE-QUOTE WHO IS THE BUFF ?QUESTION-MARK 'SINGLE-QUOTE

  THE TRANSACTION MAY OPEN THE DOOR FOR A. M. R. CORPORATION'S AMERICAN AIRLINES TO STRENGTHEN ITS ALREADY DOMINANT POSITION AT ITS DALLAS HUB .PERIOD

  TODAY THEY ARE OUR ALLIES ,COMMA TOMORROW THEY CAN ABANDON US TO OUR FATE .PERIOD

- 13% WER

  SAYS TASTE DYER ,COMMA THEY WINTHROP GROUP MANAGING DIRECTOR :COLON "DOUBLE-QUOTE MY FIRST THOUGHT IS TO GO TO THE FACTORY AND SAY ,COMMA 'SINGLE-QUOTE WHO'S THE BOSS ?QUESTION-MARK 'SINGLE-QUOTE

  THE TRANSACTION MAY OPEN THE DOOR FOR A. M. R. CORPORATION'S AMERICAN AIRLINES TO STRENGTHEN ITS ALREADY DOMINANT POSITION AT ITS DALLAS HUB .PERIOD

  TODAY THEY ARE ALLIANCE ,COMMA TOMORROW THAT CAN INDEBTEDNESS TO OUR FATE .PERIOD

- 35% WER

  SAYS STATUS TIRE ,COMMA IN WHEN TRAPPED PROVED MANAGING DIRECTOR :COLON "DOUBLE-QUOTE MY FIRST OPT IS TO GO TO THE FACTORY IN SAKE ,COMMA 'SINGLE-QUOTE EARLIEST ABOUT ?QUESTION-MARK 'SINGLE-QUOTE

  THE TRANSACTION MAY OPEN INDOOR FOUR CAME ARE CORPORATIONS AMERICAN AIRLINES TO STRENGTHEN ITS ALREADY DOMINANT POSITION BATUS PALESTINE .PERIOD

  TODAY THEIR ALLIANCE ,COMMA TOMORROW THEY COMPETITIVENESS TO A STATE .PERIOD

What WER Differences are Perceptible?

- 15.5% or 16.9% WER?

  SAYS THE A. S. TIRE ,COMMA THEY WINTHROP GROUP MANAGING DIRECTOR EAR :COLON "DOUBLE-QUOTE MY FIRST TOP IS TO GO TO THE FACTORY AND SAY ,COMMA 'SINGLE-QUOTE WHO IS THE BUFF ?QUESTION-MARK 'SINGLE-QUOTE

  THE TRANSACTION MAY OPENING OR FOR A. M. R. CORPORATION'S AMERICAN AIRLINES TO STRENGTHEN ITS ALREADY DOMINANT POSITION THAT IS A DALLAS HUB .PERIOD

  TODAY THEY ARE NOW ALLIANCE ,COMMA TOMORROW THEY CAN DENNIS TO A FADE .PERIOD

  UPJOHN SAID IT BEGAN TESTING ROGAINE ON WOMEN IN NINETEEN EIGHTY SEVEN ,COMMA ABOUT FOUR YEARS AFTER TESTS ON BALDING MAN BEGAN .PERIOD

- 15.5% or 16.9% WER?

  SAYS BASED DIRE ,COMMA THEY WINTHROP GROUP MANAGING DIRECTOR EAR :COLON "DOUBLE-QUOTE MY FIRST THOUGHT IS TO GO TO THE FACTORY AND SAY ,COMMA 'SINGLE-QUOTE HELL IS THE BUFF ?QUESTION-MARK 'SINGLE-QUOTE

  THE TRANSACTION MAY OPEN THE DOOR FOR A. M. R. CORPORATION'S AMERICAN AIRLINES TO STRENGTHEN ITS ALREADY DOMINANT POSITION AT THIS DALLAS HUB .PERIOD

  TODAY RELIANCE ,COMMA TOMORROW THEY CAN DENNIS TO AFRAID .PERIOD

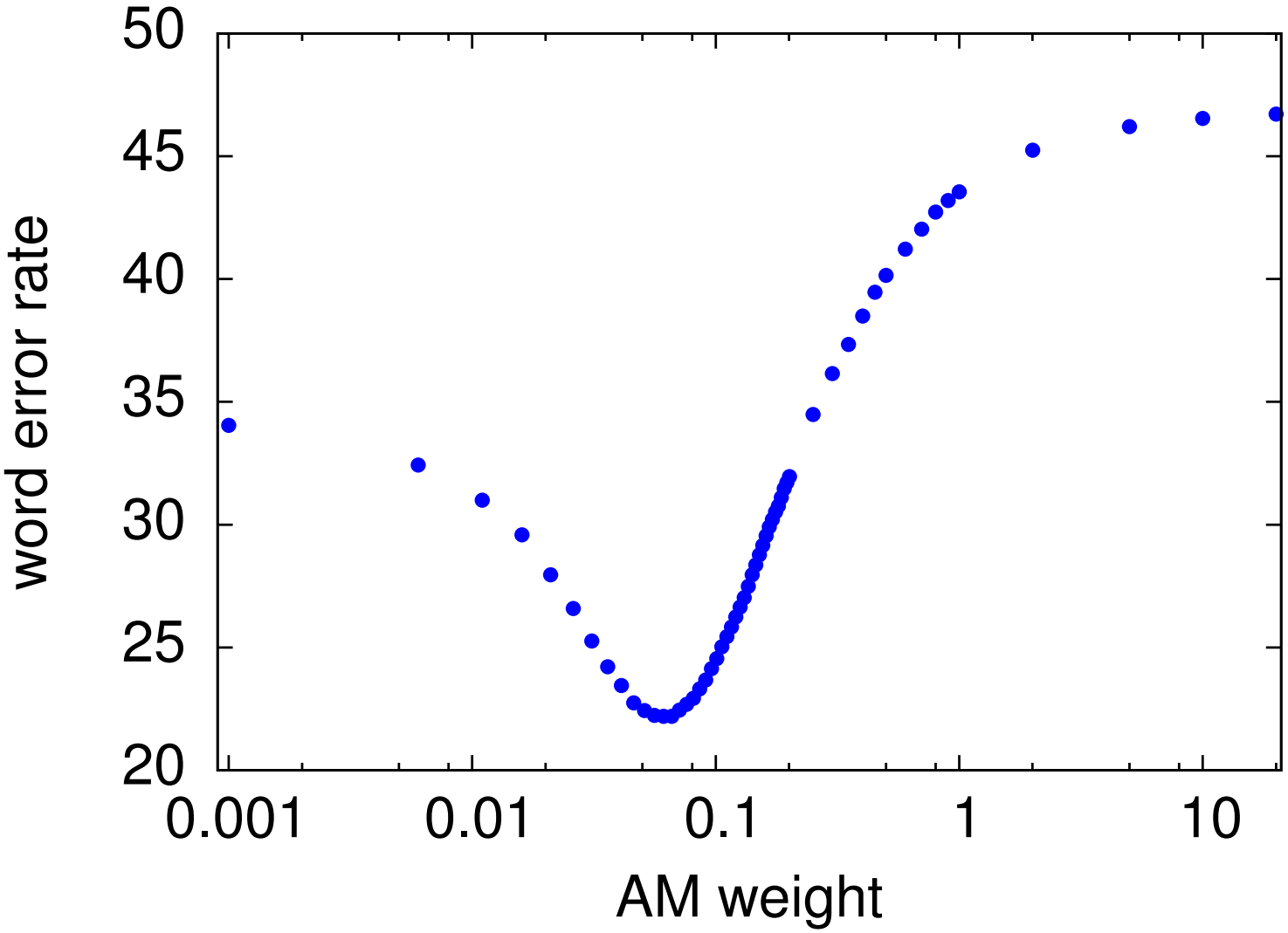
  UPJOHN SAID IT BEGAN TESTING ROGAINE ON WOMEN IN NINETEEN EIGHTY SEVEN ,COMMA ABOUT FOUR YEARS AFTER TESTS ON BALDING MAN BEGAN .PERIOD

The Acoustic Model Weight

$$\begin{aligned}(\text{answer}) &= \arg\max_{\omega} \ (\text{language model}) \times (\text{acoustic model})^{\alpha} \\ &= \arg\max_{\omega} \ P(\omega)P(\mathbf{x}|\omega)^{\alpha}\end{aligned}$$

- We smoothed/regularized the LM; what about the AM?
- AM probs are drastically overtrained!
- Fudge factor: need to tune to specific AM/LM pair.
  - α usually somewhere between 0.05 and 0.1.

Varying the Acoustic Model Weight



Evaluating Language Models

- Best way: plug into ASR system; measure WER.
  - Need ASR system; results depend on AM.
  - Painful to compute (*e.g.*, need to tune AM weight).
- Is there something cheaper that predicts WER well?

Perplexity

- Take (geometric) average word probability p_{avg}.

$$p_{\text{avg}} = \left[\prod_{i=1}^{L+1} P(w_i | w_{i-2} w_{i-1}) \right]^{\frac{1}{L+1}}$$

- Invert it: PP $= \frac{1}{p_{\text{avg}}}$.
- Interpretation: branching factor of search space.
  - *e.g.*, uniform unigram LM over k words $\Rightarrow$ PP = k.

Example: Perplexity

$$
\begin{aligned}
&P(\text{I HATE TO WAIT}) \\
&= P(\text{I}|\triangleright)P(\text{HATE}|\text{I})P(\text{TO}|\text{HATE})P(\text{WAIT}|\text{TO})P(\triangleleft|\text{WAIT}) \\
&= \frac{3234}{892669} \times \frac{45}{21891} \times \frac{40}{246} \times \frac{324}{510508} \times \frac{35}{882} = 3.05 \times 10^{-11}
\end{aligned}
$$

$$
\begin{aligned}
p_{\text{avg}} &= \left[\prod_{i=1}^{L+1} P(w_i|w_{i-1})\right]^{\frac{1}{L+1}} \\
&= (3.05 \times 10^{-11})^{\frac{1}{5}} = 0.00789 \\
\text{PP} &= \frac{1}{p_{\text{avg}}} = 126.8
\end{aligned}
$$

Perplexity: Example Values

type	domain	training data	case+ punct	PP
human[1]	biography			142
machine[2]	Brown	600MW	$\surd$	790
ASR[3]	WSJ	23MW		120

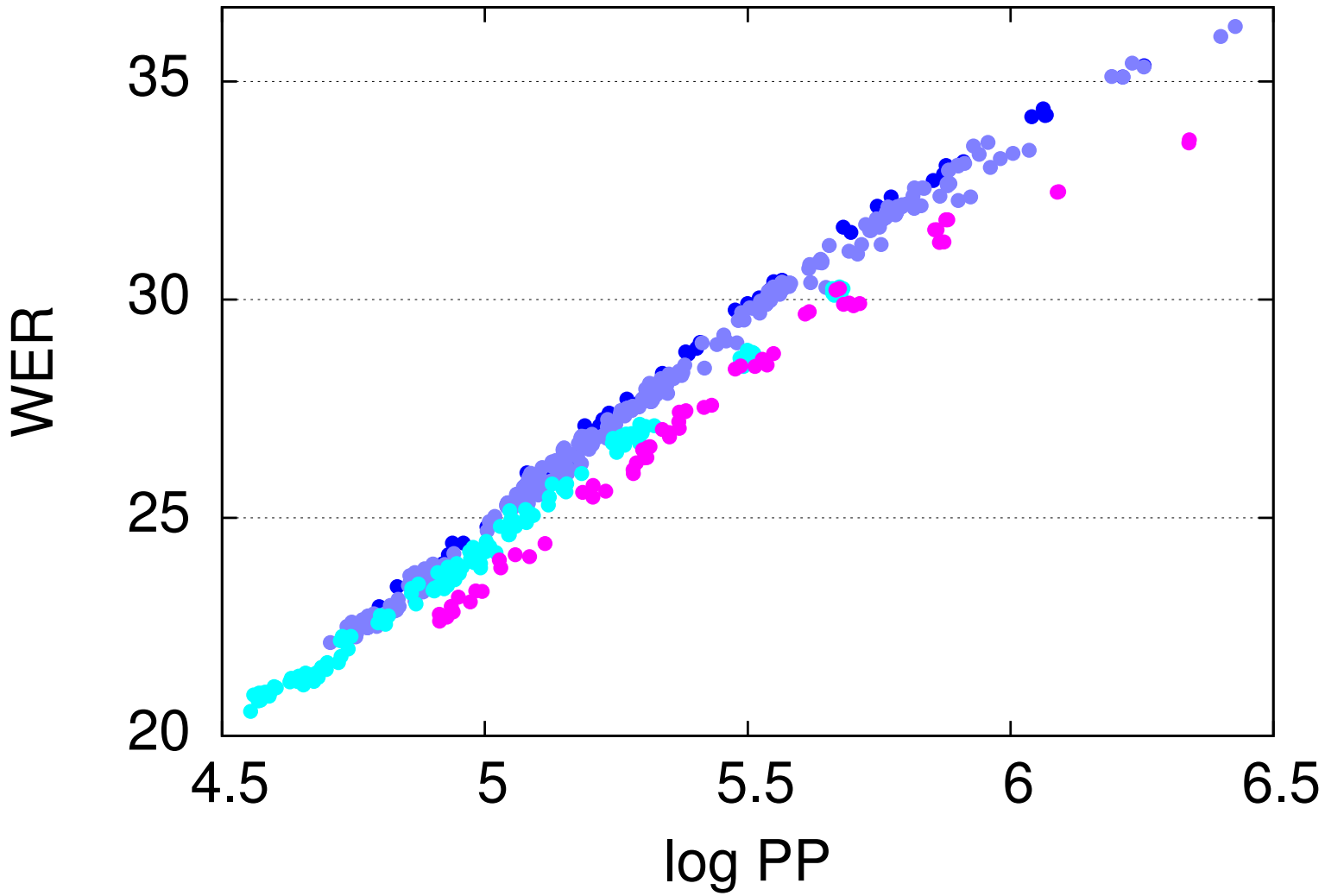
- Varies highly across domains, languages. Why?

[1] *Jefferson the Virginian*; Shannon game (Shannon, 1951).
[2] Trigram model (Brown *et al.*, 1992).
[3] Trigram model; 20kw vocabulary.

Does Perplexity Predict Word-Error Rate?

- Not *across* different LM types.
  - *e.g.*, word *n*-gram model; class *n*-gram model; . . .
- OK *within* LM type.
  - *e.g.*, vary training set; model order; pruning; . . .

Perplexity and Word-Error Rate



Recap

- Need AM weight for LM to have full effect.
- Best to evaluate LM's using WER . . .
  - But perplexity can be informative.
- What about evaluating AM's outside of ASR?
- Can you think of any problems with word error rate?
  - What do we really care about in applications?

Where Are We?

N-Gram Models

- Super simple; no linguistic knowledge.
- Workhorse of language modeling for ASR for 30+ years.
  - Used in great majority of deployed systems.
- Easy to use.
  - Fast to train; fast to run; scalable.
  - Train 4g on 1G+ words in hours.

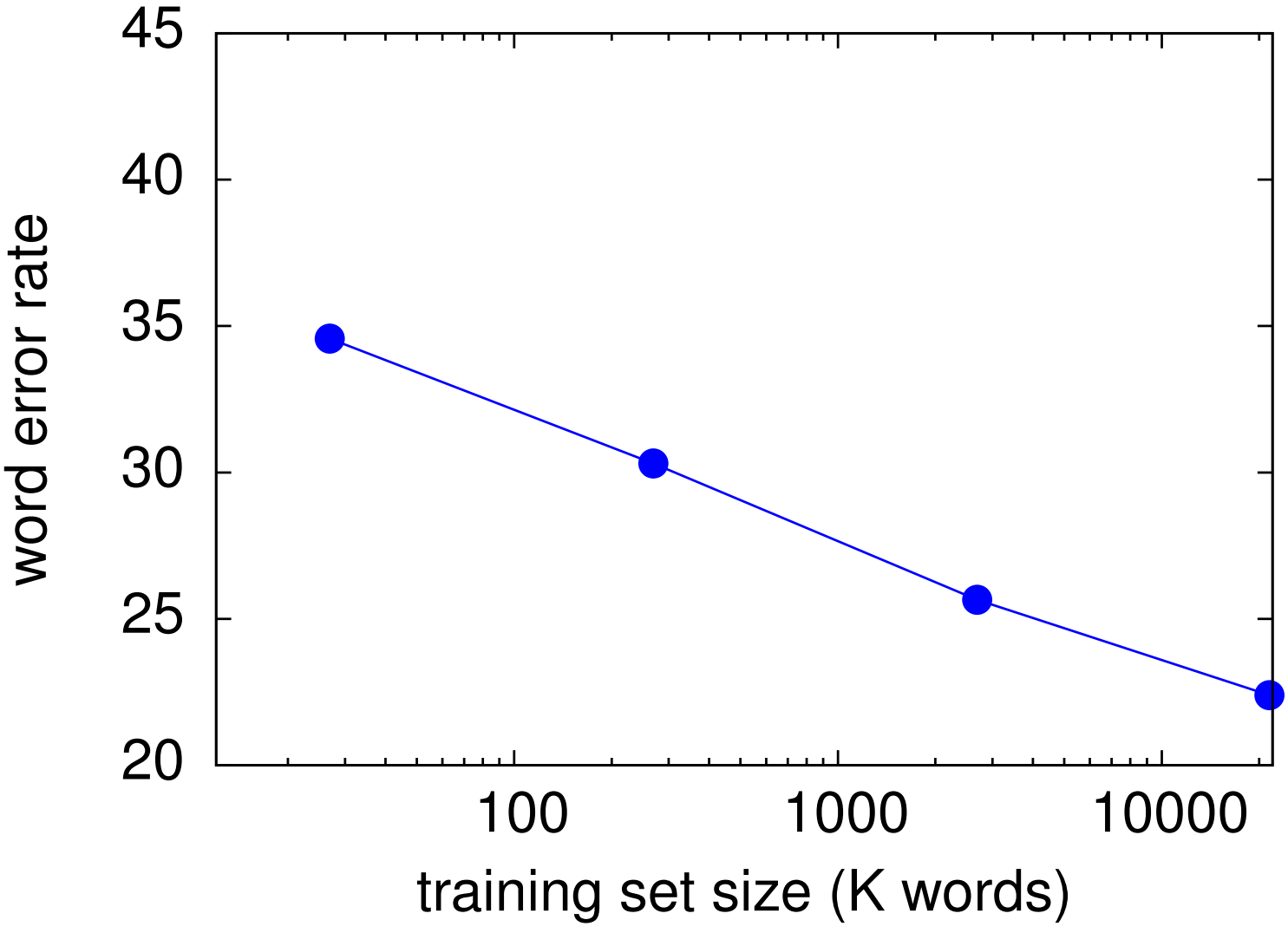
Smoothing

- Lots and lots of smoothing algorithms developed.
  - Will describe newer algorithms in later lecture.
  - Gain: $\leq$1% absolute in WER over Katz.
- Don't need to worry about models being too big!
  - No penalty from sparseness with higher n.

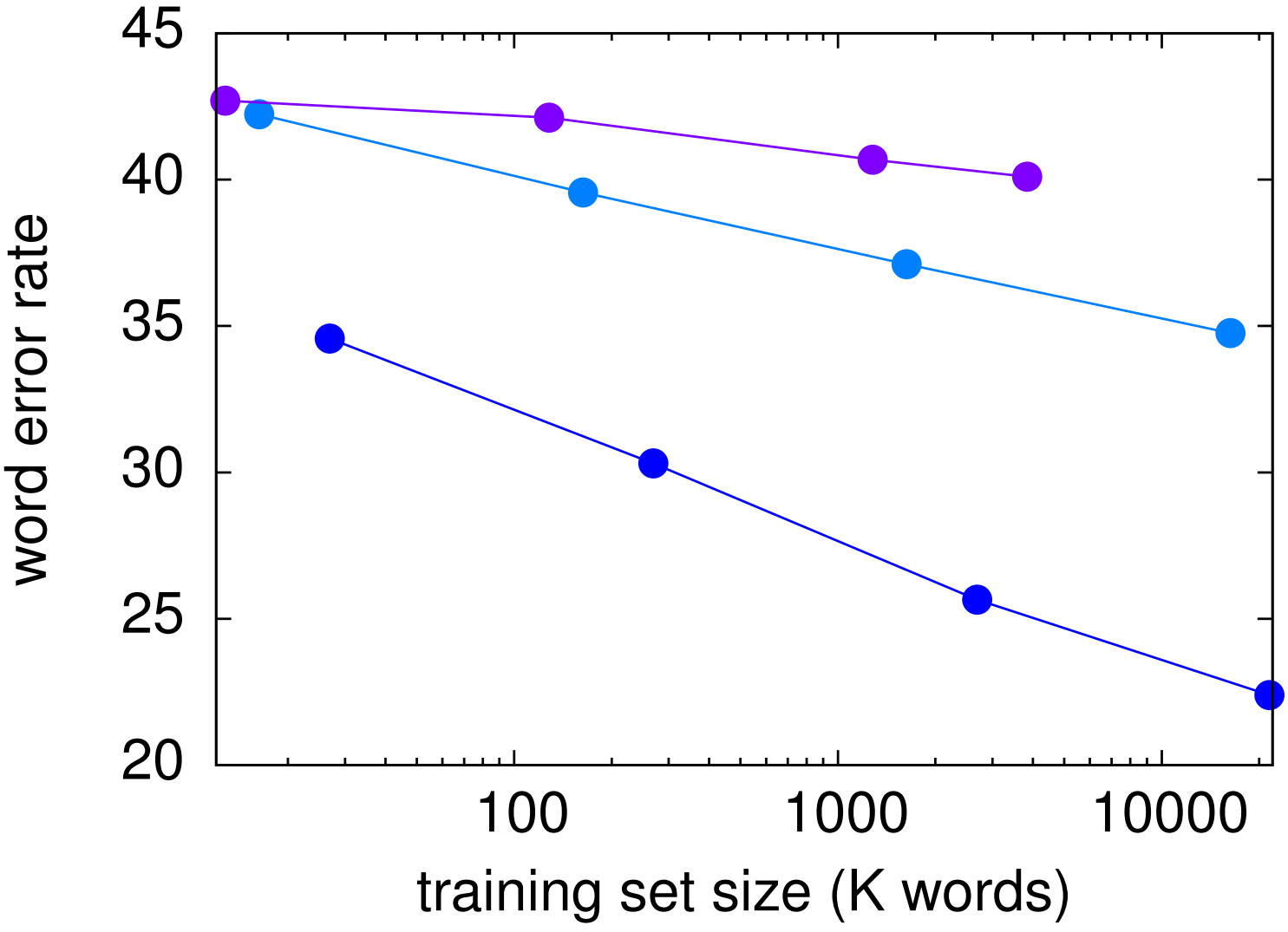
Building Language Models

- What do you need to build an LM?
- Software, *e.g.*, SRILM.
- Hyperparameters? Choose n.
- Text! $\Rightarrow$ This is what it's all about.

It's All About the Data, Part I



It's All About the Data, Part II



Case Study: Open Voice Search

- Company 1: Groogle
  - 1T searches/year (worldwide) $\Rightarrow$ $>$100GW/year (U.S.)
- Company 2: Company Not Named Wahoo or Nicrosoft.
  - Has access to only public data (10MW).
- How much better will Groogle be in WER?

Demo: Domain Mismatch

Work Smarter?

- To be continued . . .

References

- C.E. Shannon, “Prediction and Entropy of Printed English”, Bell Systems Technical Journal, vol. 30, pp. 50–64, 1951.
- I.J. Good, “The Population Frequencies of Species and the Estimation of Population Parameters”, Biometrika, vol. 40, no. 3 and 4, pp. 237–264, 1953.
- S.M. Katz, “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer”, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 35, no. 3, pp. 400–401, 1987.
- P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, J.C. Lai, R.L. Mercer, “An Estimate of an Upper Bound for the Entropy of English”, Computational Linguistics, vol. 18, no. 1, pp. 31–40, 1992.