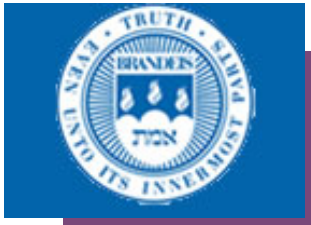
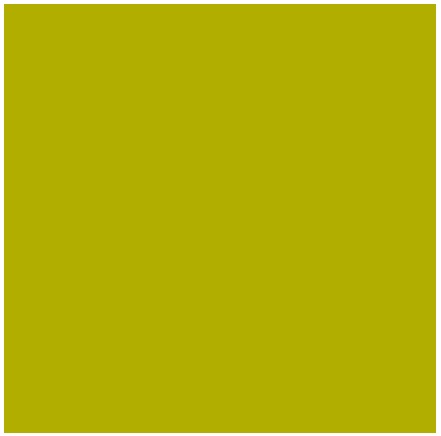
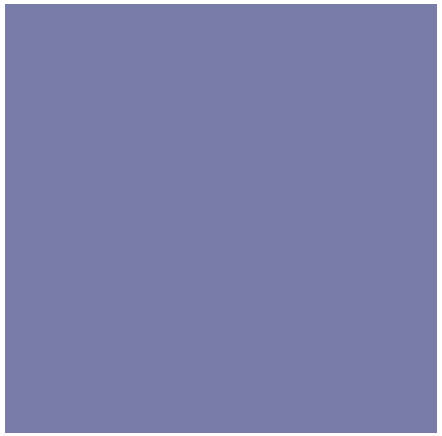
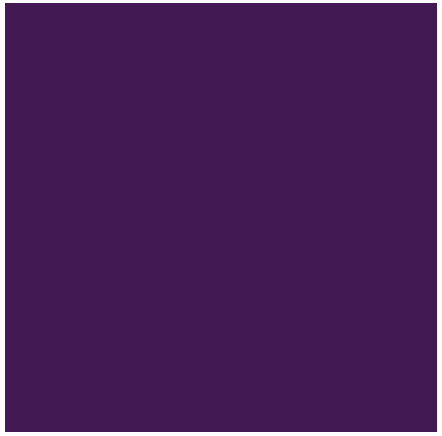








September16, 2016
Professor Meteer

CS 136a Lecture 8 Yet More Language Modeling

Thanks to Dan Jurafsky & Josh Goodman for many of these slides π

Overview (from Microsoft Tutorial)

- Caching
- Skipping
- Clustering
- Sentence-mixture models
- Structured language models
- Tools
- More on the author, Josh Goodman
  http://research.microsoft.com/en-us/um/people/joshuago/icmldescription.htm

Caching

- If you say something, you are likely to say it again later.

$$P(z \mid \text{history} = \lambda P_{smooth}(z \mid xy) + (1 - \lambda) P_{cache}(z \mid \text{history})$$

- Interpolate trigram with cache
- Trigram caches get almost twice the improvement as unigram caches

$$P_{cache}(z \mid \text{history}) = \frac{C(z \in \text{history})}{\text{length(history)}}$$

Caching: Variations

- N-gram caches:

$$P_{cache}(z \mid history) = \frac{C(xyz \in history)}{C(xy \in history)}$$

- Conditional n-gram cache: use n-gram cache only if $xy \in history$
- Remove function-words from cache, like “the”, “to”

+ Skipping

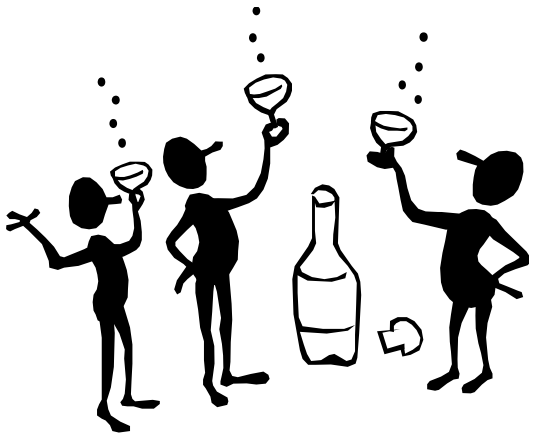
- Capturing phrasal elements
  - Show John a good time → Show XXX a good time
- Standard 5 gram: P(z|…rstuvwxy) ≈ P(z|vwxy)
- Why not P(z|v_xy) – "skipping" n-gram – skips value of 3-back word
- Example: "P(time|show John a good)" ->

  P(time | show ____ a good)
- P(…rstuvwxy) ≈

  $\lambda P(z|vwxy) + \mu P(z|vw\_y) + (1-\lambda-\mu)P(z|v\_xy)$

+ Clustering

- CLUSTERING = CLASSES (same thing)
- What is P(“Tuesday | party on”)
- Similar to P(“Monday | party on”)
- Similar to P(“Tuesday | celebration on”)
- Put words in clusters:
  - WEEKDAY = Sunday, Monday, Tuesday, …
  - EVENT=party, celebration, birthday, …



Clustering overview

- Major topic, useful in many fields
- Kinds of clustering
  - Predictive clustering
  - Conditional clustering
  - IBM-style clustering
- How to get clusters
  - Be clever or it takes forever!

Predictive clustering

- Let “z” be a word, “Z” be its cluster
- One cluster per word: hard clustering
  - WEEKDAY = Sunday, Monday, Tuesday, …
  - MONTH = January, February, April, May, June, …
- $P(z|xy) = P(Z|xy) \times P(z|xyZ)$
- P(Tuesday | party on) = P(WEEKDAY | party on) × P(Tuesday | party on WEEKDAY)
- $P_{smooth}(z|xy) \approx P_{smooth}(Z|xy) \times P_{smooth}(z|xyZ)$

Predictive clustering example

- Find P(Tuesday | party on)

  P_{smooth} (WEEKDAY | party on) ×

  P_{smooth} (Tuesday | party on WEEKDAY)

  C(party on Tuesday) = 0

  C(party on Wednesday) = 10

  C(arriving on Tuesday) = 10

  C(on Tuesday) = 100

P_{smooth} (WEEKDAY | party on) is high

P_{smooth} (Tuesday | party on WEEKDAY) backs off to P_{smooth} (Tuesday | on WEEKDAY)

+ Conditional clustering

Condition off classes in the context

$P(z|xy) = P(z|xXyY)$

P(Tuesday | party on) =

P(Tuesday | party EVENT on PREPOSITION)

$P_{smooth}(z|xy) \approx P_{smooth}(z|xXyY)$

λP_{ML} (Tuesday | party EVENT on PREPOSITION) +

μP_{ML} (Tuesday | EVENT on PREPOSITION) +

δP_{ML} (Tuesday | on PREPOSITION) +

$\gamma_{ML} P$(Tuesday | PREPOSITION) +

$(1- \lambda - \mu - \delta - \gamma) P_{ML}$ (Tuesday)

+ Conditional clustering example

Eliminating redundancy

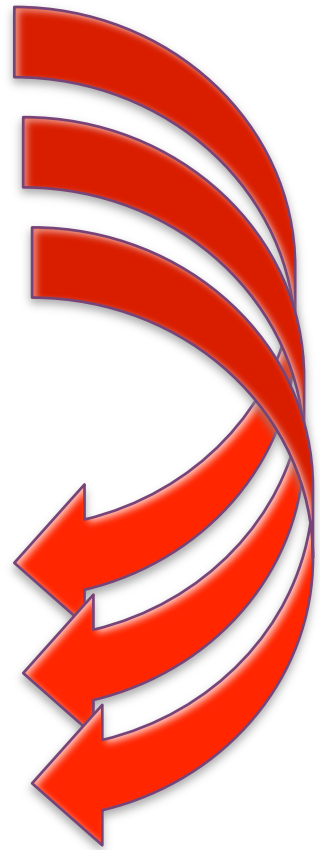
λP (Tuesday | party EVENT on PREPOSITION) +

μ P(Tuesday | EVENT on PREPOSITION) +

δP(Tuesday | on PREPOSITION) +

γP(Tuesday | PREPOSITION) +

(1- λ - μ - δ- γ) P(Tuesday



λP (Tuesday | party on) +

μ P(Tuesday | EVENT on) +

δP(Tuesday | on) +

γP(Tuesday | PREPOSITION) +

(1- λ - μ - δ- γ) P(Tuesday) =

Combined clustering

- $P(z|xy) \approx P_{smooth}(Z|xXyY) \times P_{smooth}(z|xXyYZ)$

P(Tuesday| party on) ≈
P_{smooth}(WEEKDAY | party EVENT on PREPOSITION) ×
P_{smooth}(Tuesday | party EVENT on PREPOSITION WEEKDAY)

- Much larger than unclustered, somewhat lower perplexity.

IBM Clustering

$P(z|xy) \approx P_{smooth}(Z|XY) \times P(z|Z)$

P(WEEKDAY|EVENT PREPOSITION)

$\times$ P(Tuesday | WEEKDAY)

- Small, very smooth, mediocre perplexity

$P(z|xy) \approx$

$$\lambda P_{smooth}(z|xy) + (1-\lambda)P_{smooth}(Z|XY) \times P(z|Z)$$

- Bigger, better than no clusters, better than combined clustering.
- Improvement: use P(z|XYZ) instead of P(z|Z)

Clustering by Position

- “A” and “AN”: same cluster or different cluster?
- Same cluster for predictive clustering
- Different clusters for conditional clustering
- Small improvement by using different clusters for conditional and predictive

Clustering: how to get them

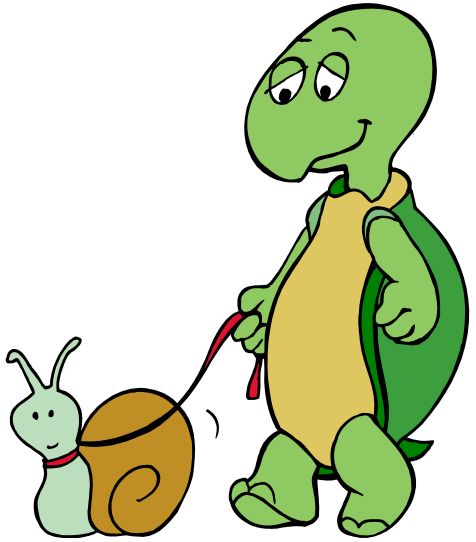
- Build them by hand
  - Works ok when almost no data
- Part of Speech (POS) tags
  - Tends not to work as well as automatic
- Automatic Clustering
  - Swap words between clusters to minimize perplexity

Clustering: automatic

- Minimize perplexity of P(z|Y)
  Mathematical tricks speed it up

Use top-down splitting,

not bottom up merging!



Two actual WSJ classes

- MONDAYS
- FRIDAYS
- THURSDAY
- MONDAY
- EURODOLLARS
- SATURDAY
- WEDNESDAY
- FRIDAY
- TENTERHOOKS
- TUESDAY
- SUNDAY
- CONDITION

- PARTY
- FESCO
- CULT
- NILSON
- PETA
- CAMPAIGN
- WESTPAC
- FORCE
- CONRAN
- DEPARTMENT
- PENH
- GUILD

Sentence Mixture Models

- Lots of different sentence types:
  - Numbers (The Dow rose one hundred seventy three points)
  - Quotations (Officials said “quote we deny all wrong doing ”quote)
  - Mergers (AOL and Time Warner, in an attempt to control the media and the internet, will merge)
- Model each sentence type separately

Sentence Mixture Models

- Roll a die to pick sentence type, s_k

with probability λ_k

- Probability of sentence, given s_k

$$\prod_{i=1}^{n} P(w_i \mid w_{i-2} w_{i-1} s_k)$$

- Probability of sentence across types:

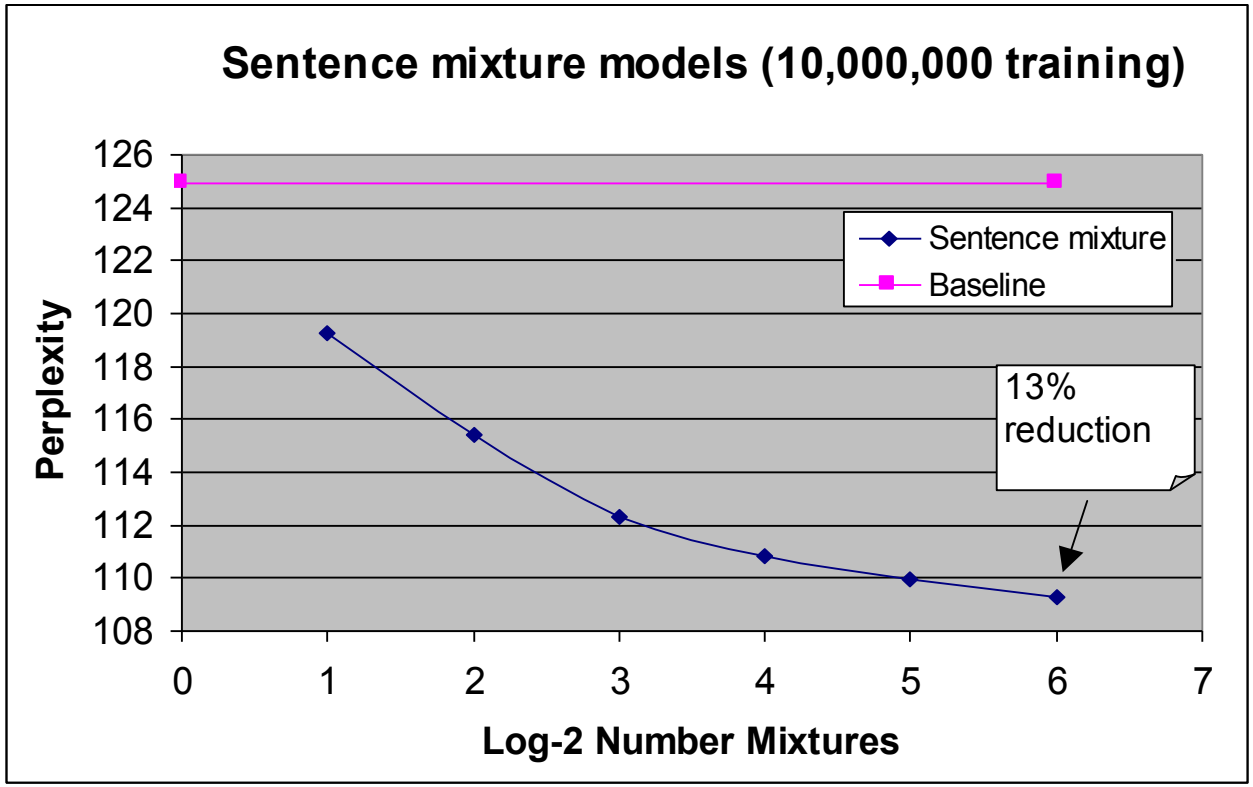
$$\sum_{k=1}^{m} \lambda_k \prod_{i=1}^{n} P(w_i \mid w_{i-2} w_{i-1} s_k)$$

Sentence Model Smoothing

- Each topic model is smoothed with overall model.
- Sentence mixture model is smoothed with overall model (sentence type 0).

$$\sum_{k=0}^{m} \lambda_k \prod_{i=1}^{n} \left[\begin{array}{l} \mu_k P(w_i \mid w_{i-2} w_{i-1} s_k) \\ + (1 - \mu_k) P(w_i \mid w_{i-2} w_{i-1}) \end{array} \right]$$

+ Sentence Mixture Results



+ Sentence Clustering

- Same algorithm as word clustering
- Assign each sentence to a type, s_k
- Minimize perplexity of $P(z|s_k)$ instead of $P(z|Y)$

Topic Examples - 0 (Mergers and acquisitions)

- JOHN BLAIR & COMPANY IS CLOSE TO AN AGREEMENT TO SELL ITS T. V. STATION ADVERTISING REPRESENTATION OPERATION AND PROGRAM PRODUCTION UNIT TO AN INVESTOR GROUP LED BY JAMES H. ROSENFIELD ,COMMA A FORMER C. B. S. INCORPORATED EXECUTIVE ,COMMA INDUSTRY SOURCES SAID .PERIOD
- INDUSTRY SOURCES PUT THE VALUE OF THE PROPOSED ACQUISITION AT MORE THAN ONE HUNDRED MILLION DOLLARS .PERIOD
- JOHN BLAIR WAS ACQUIRED LAST YEAR BY RELIANCE CAPITAL GROUP INCORPORATED ,COMMA WHICH HAS BEEN DIVESTING ITSELF OF JOHN BLAIR'S MAJOR ASSETS .PERIOD
- JOHN BLAIR REPRESENTS ABOUT ONE HUNDRED THIRTY LOCAL TELEVISION STATIONS IN THE PLACEMENT OF NATIONAL AND OTHER ADVERTISING .PERIOD
- MR. ROSENFIELD STEPPED DOWN AS A SENIOR EXECUTIVE VICE PRESIDENT OF C. B. S. BROADCASTING IN DECEMBER NINETEEN EIGHTY FIVE UNDER A C. B. S. EARLY RETIREMENT PROGRAM .PERIOD

Topic Examples - 1 (production, promotions, commas)

- MR. DION ,COMMA EXPLAINING THE RECENT INCREASE IN THE STOCK PRICE ,COMMA SAID ,COMMA "DOUBLE-QUOTE OBVIOUSLY ,COMMA IT WOULD BE VERY ATTRACTIVE TO OUR COMPANY TO WORK WITH THESE PEOPLE .PERIOD
- BOTH MR. BRONFMAN AND MR. SIMON WILL REPORT TO DAVID G. SACKS ,COMMA PRESIDENT AND CHIEF OPERATING OFFICER OF SEAGRAM .PERIOD
- JOHN A. KROL WAS NAMED GROUP VICE PRESIDENT ,COMMA AGRICULTURE PRODUCTS DEPARTMENT ,COMMA OF THIS DIVERSIFIED CHEMICALS COMPANY ,COMMA SUCCEEDING DALE E. WOLF ,COMMA WHO WILL RETIRE MAY FIRST .PERIOD
- MR. KROL WAS FORMERLY VICE PRESIDENT IN THE AGRICULTURE PRODUCTS DEPARTMENT .PERIOD

Topic Examples - 2 (Numbers)

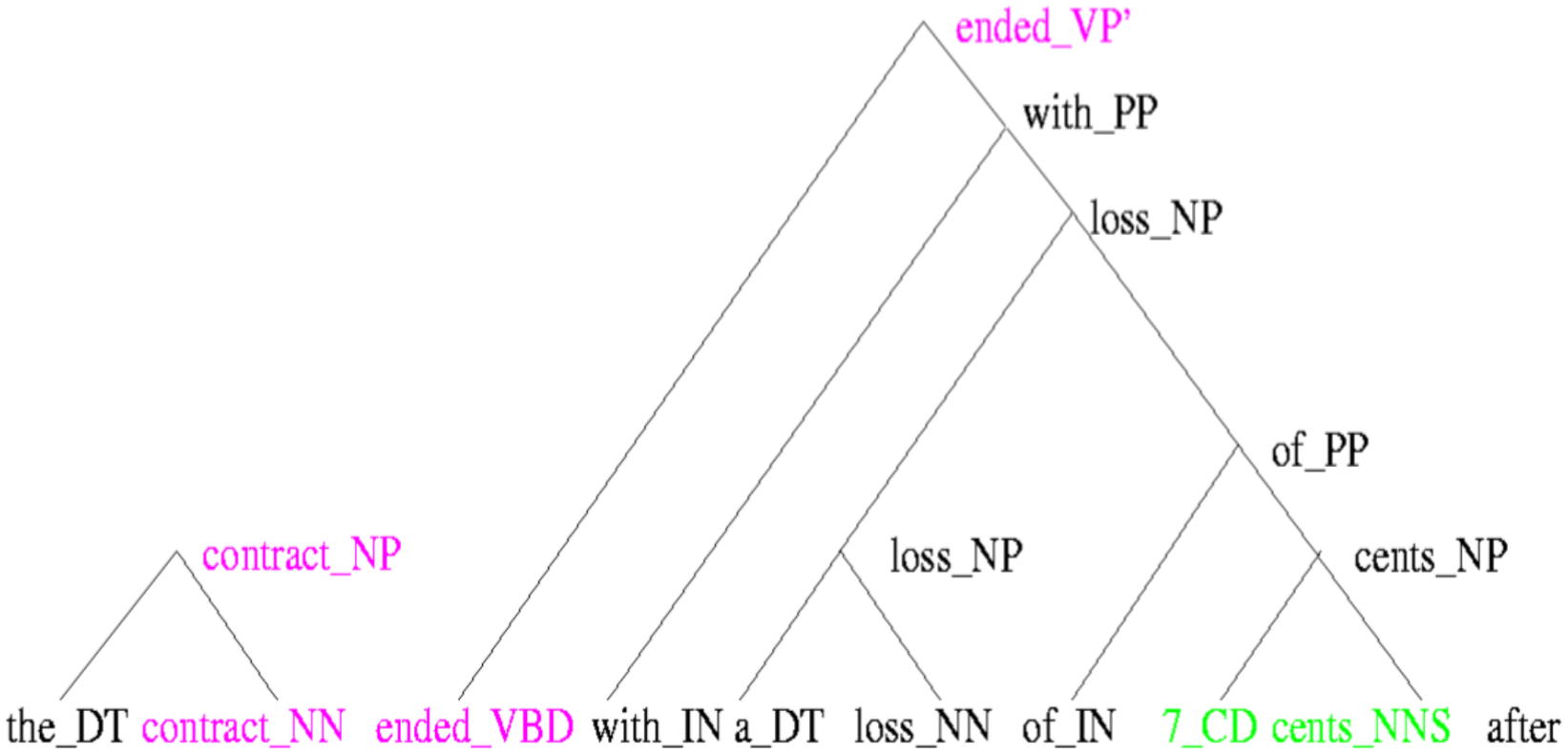
- SOUTH KOREA POSTED A SURPLUS ON ITS CURRENT ACCOUNT OF FOUR HUNDRED NINETEEN MILLION DOLLARS IN FEBRUARY ,COMMA IN CONTRAST TO A DEFICIT OF ONE HUNDRED TWELVE MILLION DOLLARS A YEAR EARLIER ,COMMA THE GOVERNMENT SAID .PERIOD
- THE CURRENT ACCOUNT COMPRISES TRADE IN GOODS AND SERVICES AND SOME UNILATERAL TRANSFERS .PERIOD
- COMMERCIAL -HYPHEN VEHICLE SALES IN ITALY ROSE ELEVEN .POINT FOUR %PERCENT IN FEBRUARY FROM A YEAR EARLIER ,COMMA TO EIGHT THOUSAND ,COMMA EIGHT HUNDRED FORTY EIGHT UNITS ,COMMA ACCORDING TO PROVISIONAL FIGURES FROM THE ITALIAN ASSOCIATION OF AUTO MAKERS .PERIOD
- INDUSTRIAL PRODUCTION IN ITALY DECLINED THREE .POINT FOUR %PERCENT IN JANUARY FROM A YEAR EARLIER ,COMMA THE GOVERNMENT SAID .PERIOD

Topic Examples – 3 (quotations)

- NEITHER MR. ROSENFIELD NOR OFFICIALS OF JOHN BLAIR COULD BE REACHED FOR COMMENT .PERIOD
- THE AGENCY SAID THERE IS "DOUBLE-QUOTE SOME INDICATION OF AN UPTURN "DOUBLE-QUOTE IN THE RECENT IRREGULAR PATTERN OF SHIPMENTS ,COMMA FOLLOWING THE GENERALLY DOWNWARD TREND RECORDED DURING THE FIRST HALF OF NINETEEN EIGHTY SIX .PERIOD
- THE COMPANY SAID IT ISN'T AWARE OF ANY TAKEOVER INTEREST .PERIOD
- THE SALE INCLUDES THE RIGHTS TO GERMAINE MONTEIL IN NORTH AND SOUTH AMERICA AND IN THE FAR EAST ,COMMA AS WELL AS THE WORLDWIDE RIGHTS TO THE DIANE VON FURSTENBERG COSMETICS AND FRAGRANCE LINES AND U. S. DISTRIBUTION RIGHTS TO LANCASTER BEAUTY PRODUCTS .PERIOD
- BUT THE COMPANY WOULDN'T ELABORATE .PERIOD

+ Structured Language Model

"The contract ended with a loss of 7 cents after"



How to get structured data?

- Use a Treebank (a collection of sentences with structure hand annotated) like Wall Street Journal, Penn Tree Bank.
- Problem: need a treebank.
- Or – use a treebank (WSJ) to train a parser; then parse new training data (e.g. Broadcast News)
- Re-estimate parameters to get lower perplexity models.

Structured Language Models

- Use structure of language to detect long distance information
- Promising results
- But: time consuming; language is right branching; 5-grams, skipping, capture similar information.

Some Experiments

- Goodman re-implemented all techniques
- Trained on 260,000,000 words of WSJ
- Optimize parameters on heldout set
- Test on separate test section
- Some combinations extremely time-consuming (days of CPU time)
  - Don't try this at home, or in anything you want to ship
- Rescored N-best lists to get results
  - Maximum possible improvement from 10% word error rate absolute to 5%

+ Overall Results: Perplexity

	Perplexity -- all-no-punc		
	Katz+	KN+	All-
perplexity	95.2	91.47	55.74
%improve &			41.45%
skip	11.16%	10.53%	5.40%
5-gram	13.89%	22.12%	22.79%
sentence	9.34%	12.97%	7.79%
cluster	-2.75%	4.56%	
cache	7.77%	7.45%	6.21%
KN	3.91%		27.80%

Overall Results: Word Accuracy

	Accuracy rates -- all-no-punc		
	Katz+	KN+	All-cache-
Accuracy	90.31	90.4	91.11
%improve &			8.26%
skip	1.03%	2.40%	1.24%
5-gram	-0.52%	2.81%	1.46%
sentence	-0.41%	-0.51%	1.35%
cluster	1.55%	3.44%	
cache	-2.99%	-1.35%	
KN	0.93%		7.54%

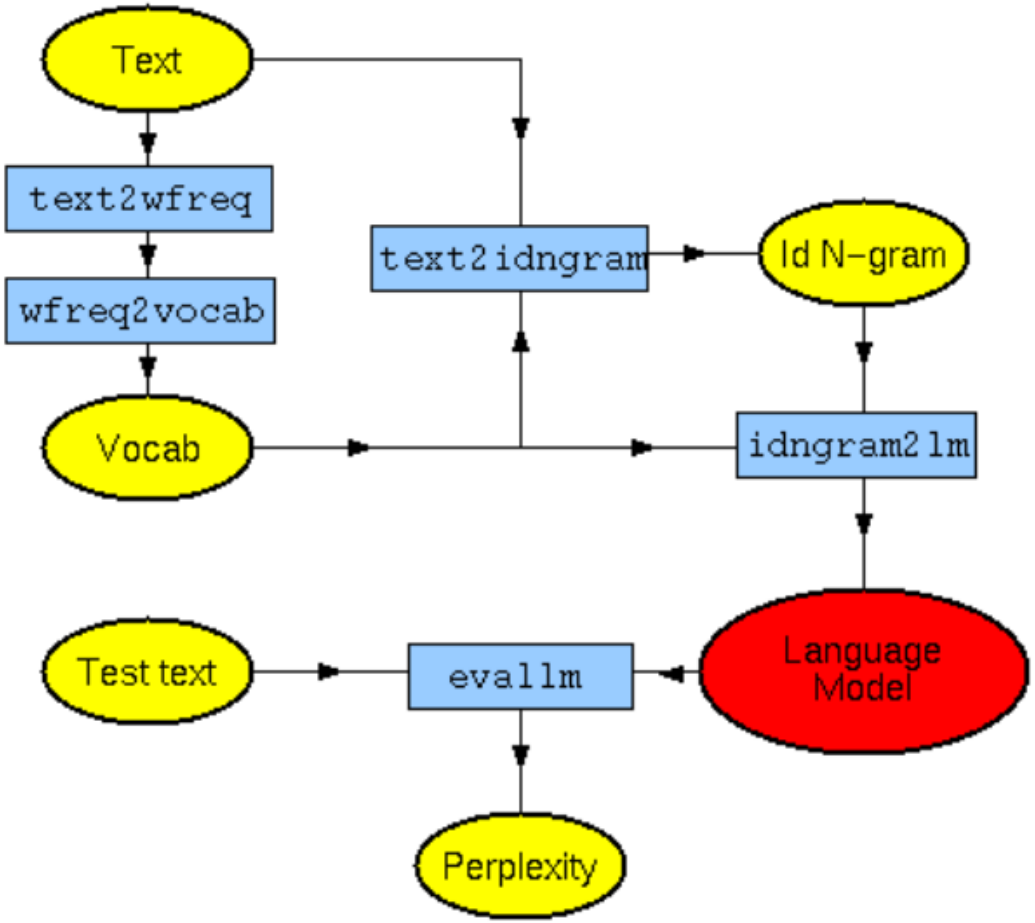
Conclusions

- Use trigram models
- Use any reasonable smoothing algorithm (Katz, Kneser-Ney)
- Use caching if you have correction information.
- Clustering, sentence mixtures, skipping not usually worth effort.

Tools: CMU Language Modeling Toolkit

- Can handle bigram, trigrams, more
- Can handle different smoothing schemes
- Many separate tools – output of one tool is input to next: easy to use
- Free for research purposes
- http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html

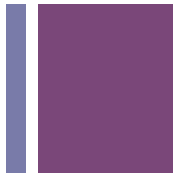
Using the CMU LM Tools



Tools: SRI Language Modeling Toolkit

- More powerful than CMU toolkit
- Can handles clusters, lattices, n-best lists, hidden tags
- Free for research use
- http://www.speech.sri.com/projects/srilm

IRSTLM



- (put in the link)
- Looks like its mostly addressing the problem of really huge LMs

Thanks to Dan Jurafsky for these slides

+ Reality: The LM is only a good as the data

- Text normalization
  - What about "$3,100,000" → convert to "Three million one hundred thousand dollars", etc.
  - Need to do this for dates, numbers, maybe abbreviations.
- Some text-normalization tools come with Wall Street Journal corpus, from LDC (Linguistic Data Consortium)
- Not much available
- Write your own (use Perl!)