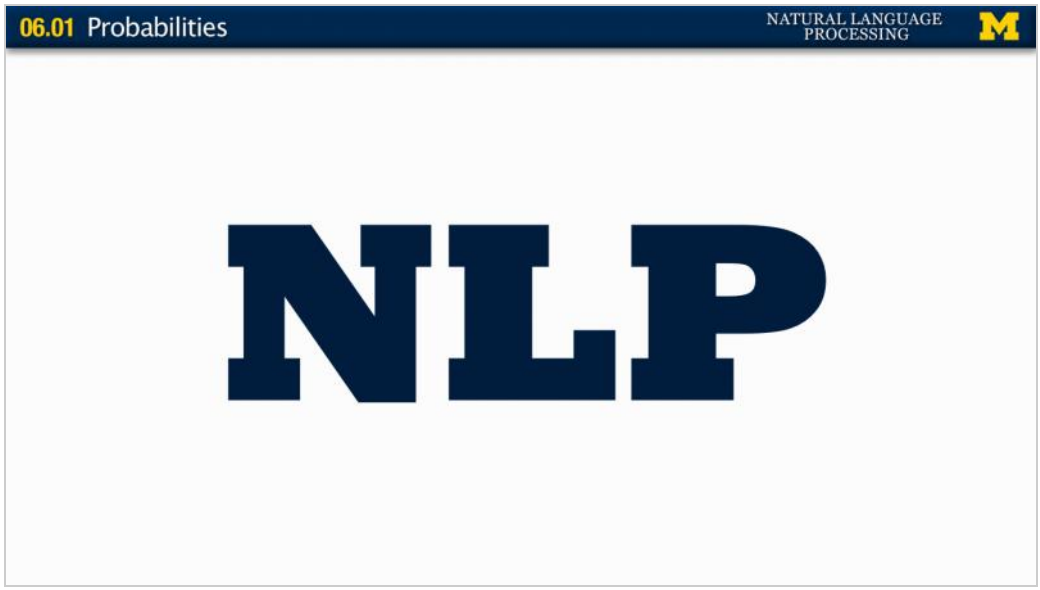
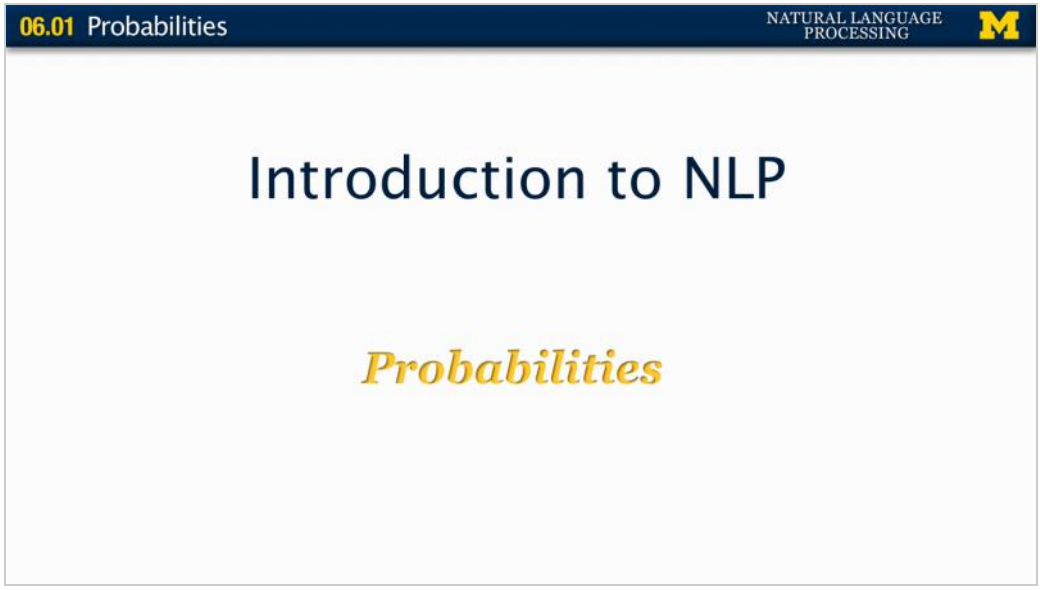
Week 6

2016年7月21日 星期四　　下午1:05

6.1 Probabilistic Language Modeling





Probabilistic Reasoning

- Very important for language processing
- Example in speech recognition:
  - "recognize speech" vs "wreck a nice beach"
- Example in machine translation:
  - "l'avocat general": "the attorney general" vs. "the general avocado"
- Probabilities make it possible to combine evidence from multiple sources in a systematic way.

Probabilities

- Probability theory
  - predicting how likely it is that something will happen
- Experiment (trial)
  - e.g., throwing a coin
- Possible outcomes
  - heads or tails
- Sample spaces
  - discrete or continuous
- Events
  - Ω is the certain event
  - ∅ is the impossible event
  - event space – all possible events

certain — either head or tail

Probabilities

- Probabilities
  - numbers between 0 and 1
- Probability distribution
  - distributes a probability mass of 1 throughout the sample space Ω.
- Example:
  - A fair coin is tossed three times.
  - What is the probability of 3 heads?
  - What is the probability of 2 heads?

Meaning of Probabilities

- Frequentist
  - I threw the coin 10 times and it turned up heads 5 times
- Subjective
  - I am willing to bet 50 cents on heads

and
— be sure that
I'm not losing money
over the long run

Properties of Probabilities

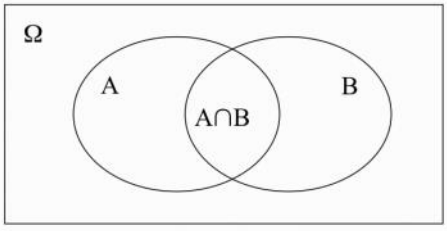
- $p(\varnothing) = 0$
- P(certain event)=1
- $p(X) \leq p(Y)$, if $X \subseteq Y$
- $p(X \cup Y) = p(X) + p(Y)$, if $X \cap Y = \varnothing$

ex: throw dice
$X = \{1,2,3,4\}$ $Y = \{1,2,3,4,5\}$.

Conditional Probability

- Prior and posterior probability
- Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



Conditional Probability

- Six-sided fair die
  - P(D even)=?
  - P(D>=4)=?
  - P(D even|D>=4)=?
  - P(D odd|D>=4)=?
- Multiple conditions
  - P(D odd|D>=4, D<=5)=?

$\frac{1}{2}$

$\frac{1}{2}$

$\frac{P(D \text{ even}, D \geq 4)}{P(D \geq 4)} = \frac{\frac{1}{3}}{\frac{1}{2}} = 2/3$

Conditional Probability

- Six-sided fair die
  - P(D even)=3/6=1/2
  - P(D>=4)=3/6=1/2
  - P(D even|D>=4)=2/3
  - P(D odd|D>=4)=1/3
- Multiple conditions
  - P(D odd|D>=4, D<=5)=1/2

$\frac{P(\text{odd}, D \geq 4, D \leq 5)}{P(D \geq 4, D \leq 5)} = \frac{\frac{1}{6}}{\frac{2}{6}} = \frac{1}{2}$

Chain rule

The Chain Rule ★

- $P(w_1,w_2,w_3 \ldots w_n) = ?$
- Using the chain rule:
  - $P(w_1,w_2,w_3 \ldots w_n) = P(w_1)\ P(w_2|w_1)\ P(w_3|w_1,w_2) \ldots P(w_n|w_1,w_2 \ldots w_{n-1})$
- This rule is used in many ways in statistical NLP, more specifically in Markov Models

★ used many times. specifically in Markov Models

Independence

- Two events are independent when
  - $P(A \cap B) = P(A)P(B)$
- Unless P(B)=0 this is equivalent to saying that P(A) = P(A|B)
- If two events are not independent, they are considered dependent

Independent events
$P(A \cap B) = P(A)\,P(B)$

Adding vs. Removing Constraints

- Adding constraints
  - P(walk=yes|weather=nice)
  - P(walk=yes|weather=nice,freetime=yes,crowded=yes)
  - More accurate
  - But more difficult to estimate
- Removing constraints (Backoff)
  - P(walk=yes|weather=nice,freetime=yes,crowded=yes)
  - P(walk=yes|weather=nice,freetime=yes)
  - P(walk=yes|weather=nice)
  - Note that it is *not* possible to do backoff on the left hand side of the conditional

removed

LHS
$P(A, B \mid C, D)$
$\not\Rightarrow P(A \mid C D)$

Random Variables

- Simply a function: $X: \Omega \rightarrow R^n$
- The numbers are generated by a *stochastic process* with a certain probability distribution
- Example
  - the discrete random variable X that is the sum of the faces of two randomly thrown fair dice
- Probability mass function (pmf) which gives the probability that the random variable has different numeric values: $P(x) = P(X = x) = P(A_x)$ where $A_x = \{\omega \in \Omega : X(\omega) = x\}$

Random Variables

- If a random variable X is distributed according to the pmf $p(x)$, then we write $X \sim p(x)$
- For a discrete random variable, we have

$$\sum_{i} p(x_i) = P(\Omega) = 1$$

Example

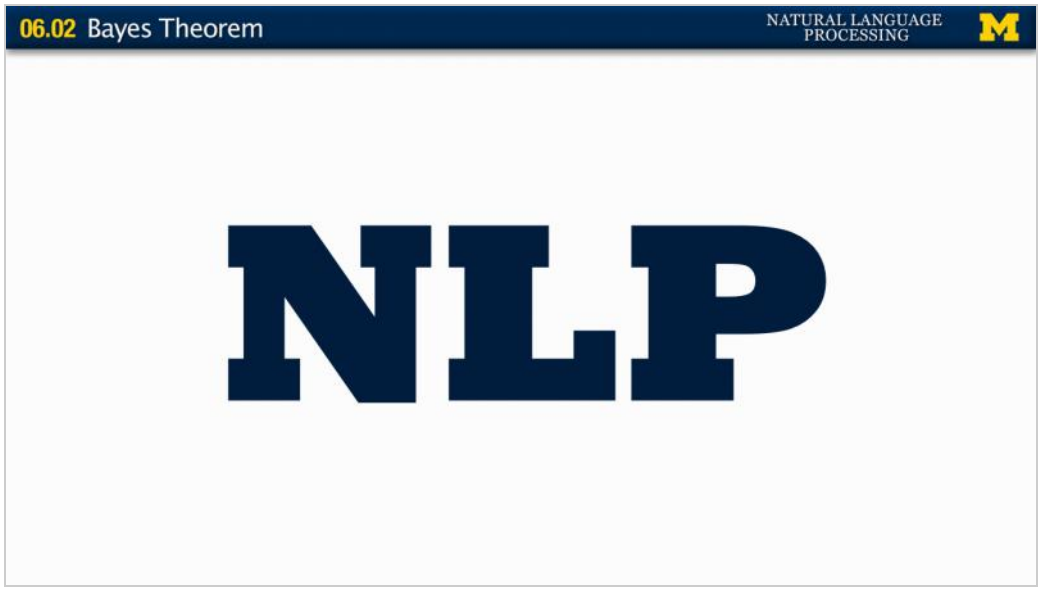
- p(1) = 1/6
- p(2) = 1/6
- etc.
- **P(D)**=?
- **P(D)** = {1/6, 1/6, 1/6, 1/6, 1/6, 1/6}
- **P(D**|odd) = {1/3, 0, 1/3, 0, 1/3, 0}

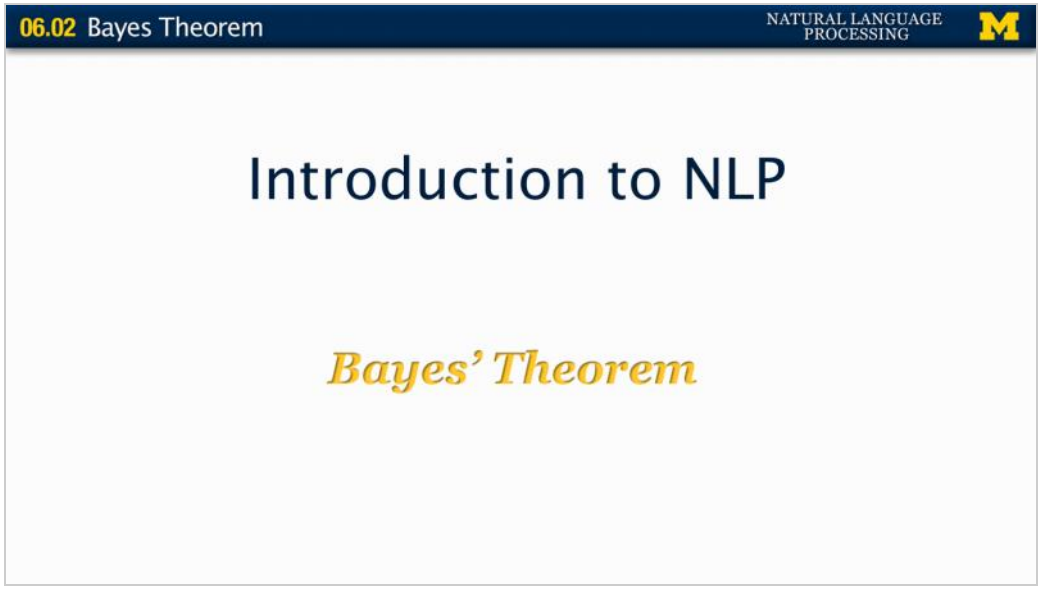
NLP

6.2

Bayes Theorem

Most important for Probabilistic NLP.





Bayes' Theorem

- Formula for joint probability
  - p(A,B) = p(B|A)p(A)
  - p(A,B) = p(A|B)p(B)
- Therefore
  - p(B|A)=p(A|B)p(B)/p(A)
- Bayes' theorem is used to calculate P(A|B) given P(B|A)

$P(A,B) = P(B|A)P(A)$

$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

Diagnostic Test

Example

- Diagnostic test
- Test accuracy
  - P(positive | ¬disease) = 0.05 – false positive
  - P(negative | disease) = 0.05 – false negative
  - so p(positive | disease) = 1–0.05 = 0.95

not sick

POS tag

Example

- Diagnostic test with errors

P(A|B)		A=TEST	
		Positive	Negative
B=DISEASE	Yes	0.95	0.05
	No	0.05	0.95

Example

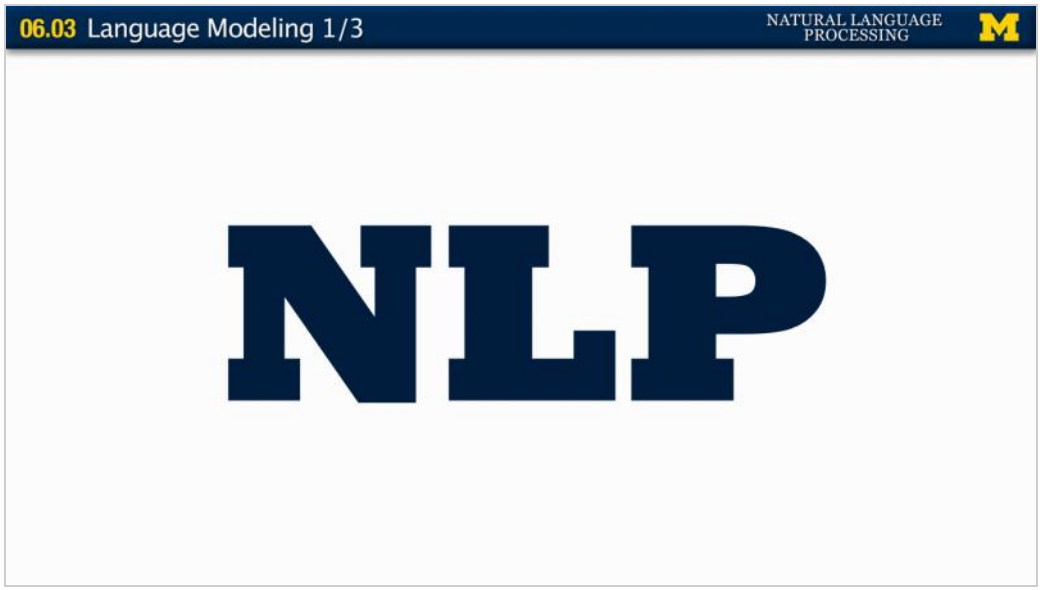
- What is p(disease | positive)?
  - P(disease|positive) = P(positive|disease)*P(disease)/P(positive)
  - P(¬disease|positive) = P(positive| ¬disease)*P(¬disease)/ P(positive)
  - P(disease|positive)/P(¬disease|positive) = ?
- We don't really care about p(positive)
  - as long as it is not zero, we can divide by it on both sides

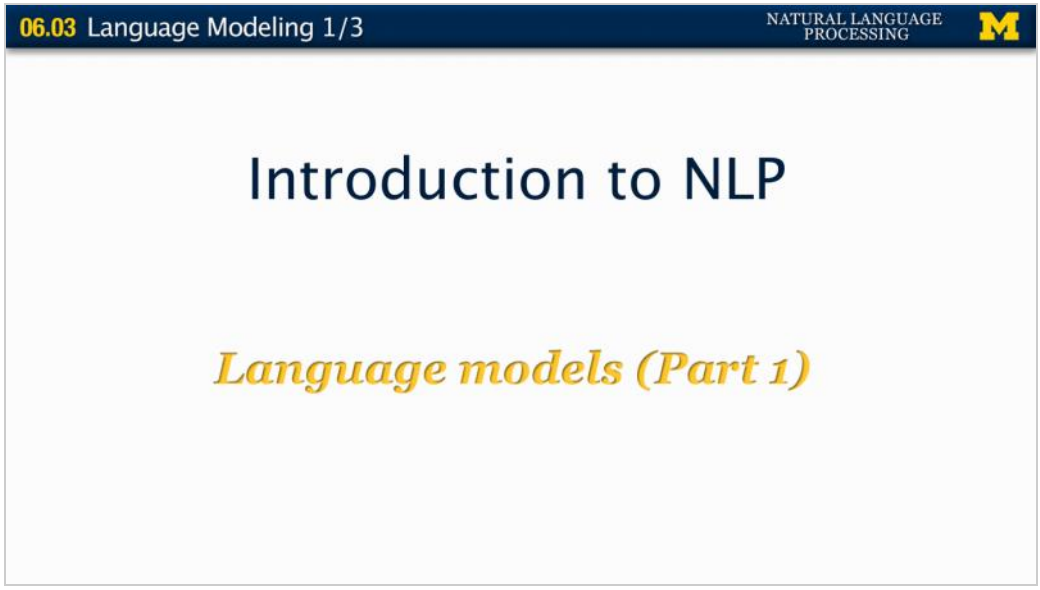
Example

- P(disease|positive)/P(¬disease|positive) = (P(positive|disease) x P(disease))/(P(positive|¬disease) x P(¬disease))
- Suppose P(disease) = 0.001
  - so P(¬disease) = 0.999
- P(disease|positive)/P(¬disease|positive) = (0.95 x 0.001)/(0.05 x 0.999) =0.019
- P(disease|positive)+P(¬disease|positive) = 1
- P(disease|positive) ≈ 0.02
- P(disease) is called the prior probability
- P(disease|positive) is called the posterior probability
- In this example the posterior is 20 times larger than the prior

NLP

6.3


06.03 Language Modeling 1/3
NATURAL LANGUAGE PROCESSING
NLP


06.03 Language Modeling 1/3
NATURAL LANGUAGE PROCESSING
Introduction to NLP
Language models (Part 1)

Probabilistic Language Models

take sentence as input and assign probability to it.

- Assign a probability to a sentence
  - $P(S) = P(w_1,w_2,w_3,...,w_n)$
- Different from deterministic methods using CFG
- The sum of the probabilities of all possible sentences must add up to 1

Predicting The Next Word

- Example
  - Let's meet in Times ... square has Pr high
  - General Electric has lost some market ... "shares" has high Pr
- Formula
  - $P(w_n|w_1,w_2,...,w_{n-1})$

Predicting the Next Word

- What word follows "your"?
  - http://norvig.com/ngrams/count_2w.txt
- your abilities 160848
  your ability 1116122
  your ablum 112926
  your academic 274761
  your acceptance 783544
  your access 492555
  your accommodation 320408
  your account 8149940
  your accounting 128409
  your accounts 257118
  your action 121057
  your actions 492448
  your activation 459379
  your active 140797
  your activities 226183
  your activity 156213
  your actual 302488
  your ad 1450485
  your address 1611337
  your admin 117943
  your ads 264771
  your advantage 242238
  your adventure 109658
  your advert 101178
  your advertisement 172783

Uses of Language Models

- Speech recognition
  - P("recognize speech") > P("wreck a nice beach")
- Text generation
  - P("three houses") > P("three house")
- Spelling correction
  - P("my cat eats fish") > P("my xat eats fish")
- Machine translation
  - P("the blue house") > P("the house blue")
- Other uses
  - OCR
  - Summarization
  - Document classification
- Usually coupled with a translation model (later)

☆ Usage of language modeling

☆ Prob. of sentence coming from English/French etc.

Probability Of A Sentence

- How to compute the probability of a sentence?
- What if the sentence is novel?
- What we need to estimate:
  - $P(S)=P(w_1,w_2,w_3 \ldots w_n)$
- Using the chain rule:
  - $P(S)= P(w_1)\ P(w_2|w_1)\ P(w_3|w_1,w_2) \ldots\ P(w_n|w_1,w_2 \ldots w_{n-1})$
- Example:
  - P("I would like the pepperoni and spinach pizza")=?

Joint prob.

N-gram Models

- Predict the probability of a word based on the words before:
  - P(square|Let's meet in Times)
- Markov assumption
  - Only look at limited history
- N-gram models
  - Unigram – no context: P(square)
  - Bigram: P(square|Times)
  - Trigram: P(square|in Times)

limited history

Random Text (Brown Corpus)

- 2-grams:

The 53-year-old Shea was no acceptable formula to help the abuse of events were a wall in 1908 , called upon his hand in Southern New Orleans , Miss Garson was named Maurice Couve De Havilland signed a privilege resolution had had happened on a tax applied to the Chisholm , the thriving systems of the `` Pride and musician , and Moscow made good team spirit of the culmination of the metal tube through the amateur , but rather than a special prosecutor . This knowledge of each member of these savings of golf course can see the 13 straight 69 . Since 1927 by Harry Truman Cleveland of railroad retirement age groups . No Vacancy " . `` I have to congressmen . The remainder of the rear bumper and on a benefit in U.S. amateur , as far as a thrill a $100 U.S. if not indicted . The state's occupation tax dollars over the newest product of the address he attended Arlington State University will pay half years .

06.03 Language Modeling 1/3 NATURAL LANGUAGE PROCESSING

Random Text (Brown Corpus)

- 3-grams:

The Fulton County Jail and `` a very strong central government of Laos that the presence of picket lines and featuring a flared skirt and lace jacket with bateau neckline and princesse skirt accented by lace appliques . Her acting began with the members of the government -- such control is necessary to build in a final exchange between Moscow and Washington last week . Of course , since the views of another one . It urged that the games are not essential to provide federal contributions to the 85-student North Carolina group to play , was addressing a meeting in the manufacture of a tax bill since most of his uncle and aunt , also was particularly struck by the reams came in from shareholders of these co-operative systems , the 9th precinct of the guiding spirits of the Armed Services Committee . Davis received 1,119 votes in Saturday's election , the executive organs of participation can hardly escape the impression that he made no attempt to get agreement among the conference's top four in rushing , he was awarded the top but that presently we're not acting as we head that way .

06.03 Language Modeling 1/3 NATURAL LANGUAGE PROCESSING

Random Text (Brown Corpus)

- 4-grams:

The broadcast said Anderson , a Seattle ex-marine and Havana businessman , and McNair , of Miami , were condemned on charges of smuggling arms to Cuban rebels . Anderson operated three Havana automobile service stations and was commander of the Havana American Legion post before it disbanded since the start of August have shown gains averaging nearly 10% above last year . That , too , in improving motorists' access to many turnpikes . The Kansas Turnpike offers an illustration . Net earnings of that road rose from 62 per cent of the prices that the avid buyers bid it up to . Dallas and North Texas is known world-wide as the manufacturing and distribution center of cotton gin machinery and equipment . The firm is design-conscious , sales-conscious , advertising-conscious . `` Hodges predicted : ' I think we should certainly follow through on it " . Rep. Henry C. Grover , who teaches history in the Houston public schools , would reduce from 24 to 12 semester hours the so-called `` blue law " controversy came in the form of a letter to Mayor Grady that plowing and salting crews should be dispatched earlier in storms and should be kept on the job .

Higher Order n-grams

- It is possible to go to 3,4,5-grams
- Longer n-grams suffer from sparseness

N-grams

- Shakespeare unigrams
  - 29,524 types, approx. 900K tokens
- Bigrams
  - 346,097 types, approx. 900K tokens
- Notice - very sparse data!

Estimation

- Can we compute the conditional probabilities directly?
  - No, because the data is sparse
- Markov assumption
  - P("musical | I would like two tickets for the") = P("musical | the")
  - or
  - P("musical | I would like two tickets for the") = P("musical | for the")

Approximate

Markov.

Approximation

Maximum Likelihood Estimates

MLE.

- Use training data
- Count how many times a given context appears in it.
- Unigram example:
  - The word "pizza" appears 700 times in a corpus of 10,000,000 words.
  - Therefore the MLE for its probability is P'("pizza") = 700/10,000,000 = 0.00007
- Bigram example:
  - The word "with" appears 1,000 times in the corpus.
  - The phrase "with spinach" appears 6 times
  - Therefor the MLE for P'(spinach|with) = 6/1,000 = 0.006
- These estimates may not be good for corpora from other genres

Example

- P("<S> I will see you on Monday</S>") =
  - P(I|<S>)
  - x P(will|I)
  - x P(see|will)
  - x P(you|see)
  - x P(on|you)
  - x P(Monday|on)
  - x P(</S>|Monday)

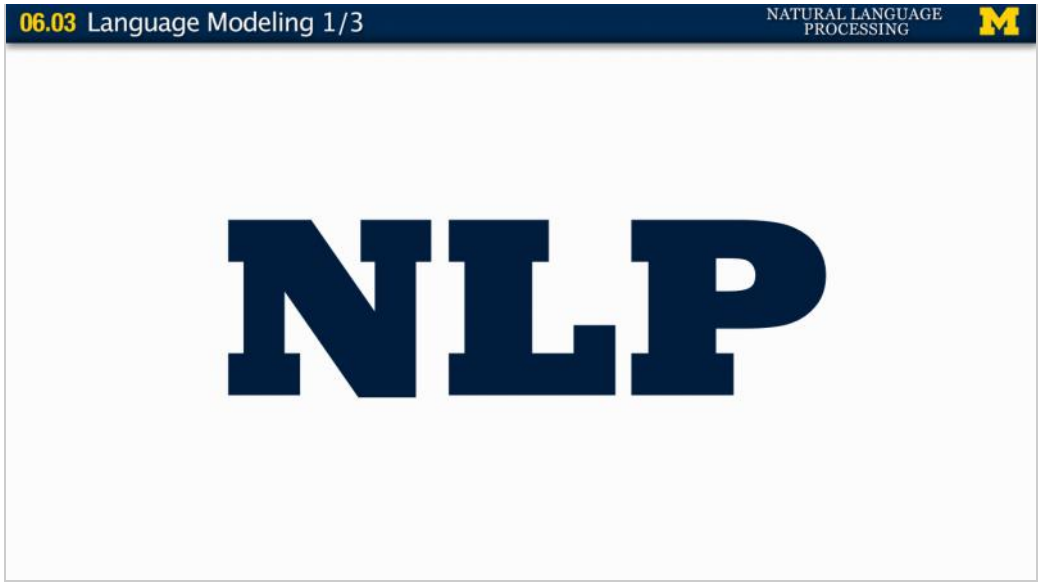
Example from Jane Austen

- P("Elizabeth looked at Darcy")
- Use maximum likelihood estimates for the n-gram probabilities
  - unigram: $P(w_i)=c(w_i)/V$
  - bigram: $P(w_i|w_{i-1}) = c(w_{i-1},w_i)/c(w_{i-1})$
- Values
  - P("Elizabeth") = 474/617091 = .000768120
  - P("looked|Elizabeth") = 5/474 = .010548523
  - P("at|looked") = 74/337 = .219584569
  - P("Darcy|at") = 3/4055 = .000739827
- Bigram probability
  - P("Elizabeth looked at Darcy") = .000000001316 = 1.3×10^{-9}
- Unigram probability
  - P("Elizabeth looked at Darcy") = 474/617091 * 337/617091 * 4055/617091 * 304/617091 = .000000000001357 = 1.3×10^{-12}
- P("looked Darcy Elizabeth at") = ?

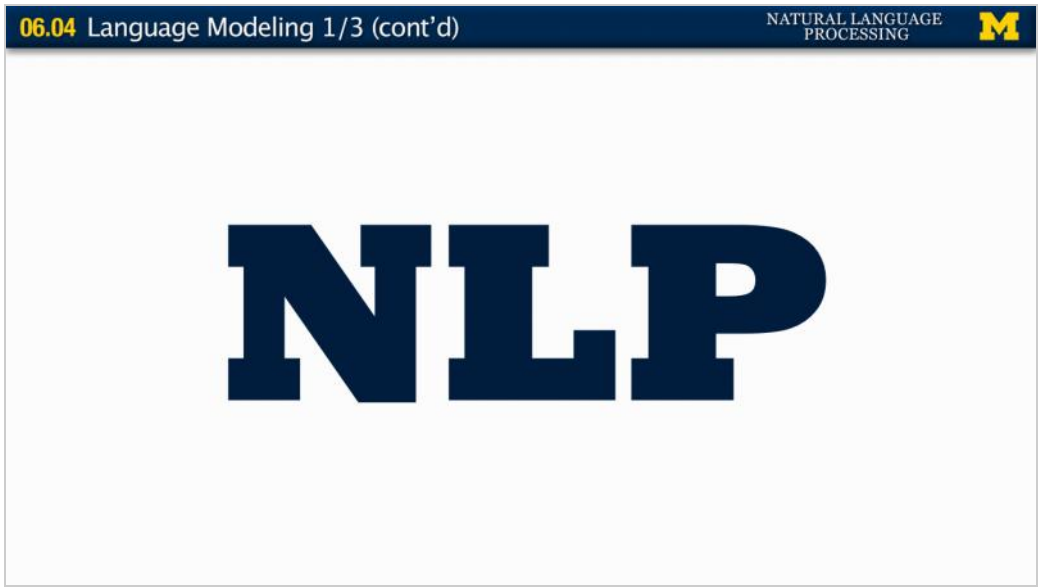
MLE

P(looked) x P(Darcy|looked) x P(Elizabeth|Darcy) x P(at|Elizabeth)

same unigram probability.


06.03 Language Modeling 1/3
NATURAL LANGUAGE PROCESSING
NLP

6.4


06.04 Language Modeling 1/3 (cont'd)
NATURAL LANGUAGE PROCESSING
NLP

N-grams And Regular Languages

- N-grams are just one way to represent weighted regular languages
- More about this in the lecture on regular languages

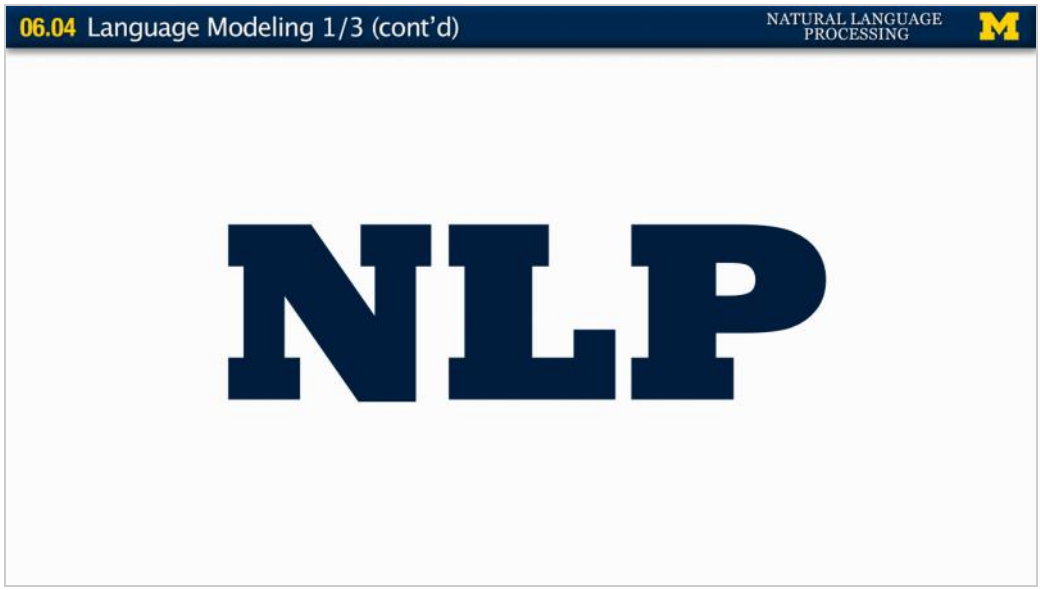
Generative Models

- Unigram: generate a word, then generate the next one, until you generate </S>.

- Bigram: generate <S>, generate a word, then generate the next one based on the previous one, etc., until you generate </S>.

Engineering Trick

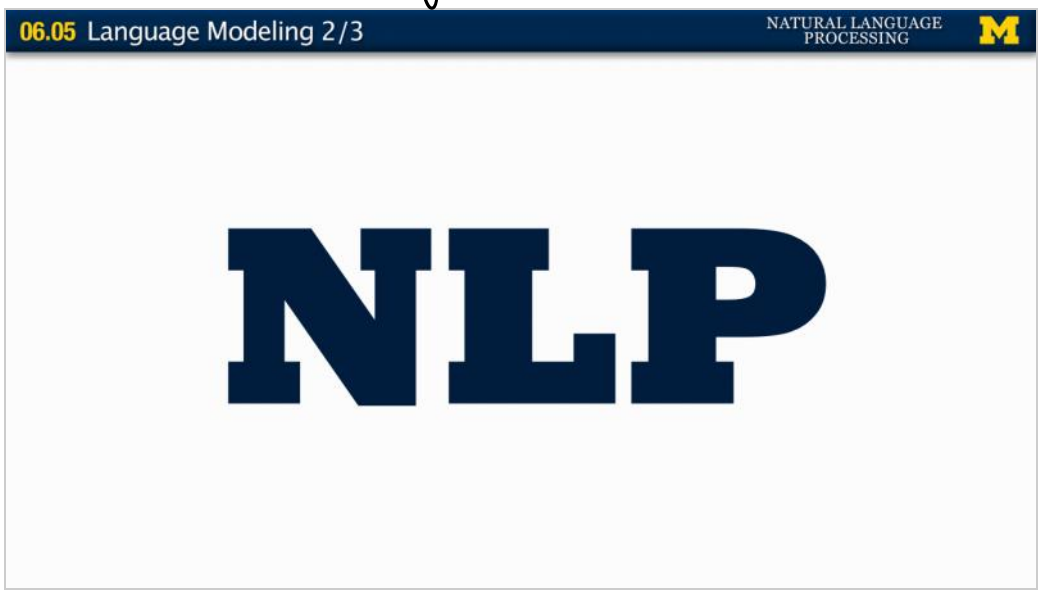
- The MLE values are often on the order of 10^{-6} or less
  - Multiplying 20 such values gives a number on the order of 10^{-120}
  - This leads to underflow
- Use (base 10) logarithms instead
  - 10^{-6} becomes -6
  - Use sums instead of products



6.5

Smoothing

Small training set
large vocab size



Introduction to NLP

Language models (Part 2)

Smoothing

- If the vocabulary size is |V|=1M
  - Too many parameters to estimate even a unigram model
  - MLE assigns values of 0 to unseen (yet not impossible) data
  - Let alone bigram or trigram models
- Smoothing (regularization)
  - Reassigning some probability mass to unseen data

if one gram not seen then the whole sentence has 0 probability.

Smoothing

- How to model novel words?
  - Or novel bigrams?
- Distributing some of the probability mass to allow for novel events
- Add-one (Laplace) smoothing:
  - Bigrams: $P(w_i|w_{i-1}) = (c(w_{i-1},w_i)+1)/(c(w_{i-1})+V)$
  - This method reassigns too much probability mass to unseen events
- Possible to do add-k instead of add-one
- Both of these don't work well in practice

novel words: have not seen in training data

Assume every word is seen one more time than it is in the data

→ reassign too much to unseen

☆ Add-1 or Add-k is not usually used because of its problem

Advanced Smoothing

- Good-Turing
  - Try to predict the probabilities of unseen events based on the probabilities of seen events
- Kneser-Ney
- Class-based n-grams

collapse words into category
⇒ predict freq. of words in this cat. based on the else in the cat

Example

- Corpus:
  - cat dog cat rabbit mouse fish fish mouse hamster hamster fish turtle tiger cat rabbit cat dog dog fox lion
- What is the probability the next item is "mouse"?
  - P_{MLE} (mouse) = 2/20
- What is the probability the next item is "elephant" or some other previously unseen animal?

not seen before.

  - Trickier
  - Is it 0/20?
  - Note that P (that the next animal is unseen) > 0
  - Therefore we need to discount the probabilities of the animals that have already been seen
  - P_{MLE} (mouse) < 2/20

Good Turing

- Actual counts c
- N_r = number of n-grams that occur exactly c times in the corpus
- N_0 = total number of n-grams in the corpus
- Revised counts c*
  - $c^* = (c+1)\, N_{c+1}/N_c$

# of n-grams that occur exactly c times in the corpus.

Example

- Corpus:
  - cat dog cat rabbit mouse fish fish mouse hamster hamster fish turtle tiger cat rabbit cat dog dog fox lion
- Counts
  - C(cat) = 4
  - C(dog) = 3
  - C(fish) = 3
  - C(mouse) = 2
  - C(rabbit) = 2
  - C(hamster) = 2
  - C(fox) = 1
  - C(turtle) = 1
  - C(tiger) = 1
  - C(lion) = 1
- N1=4, N2=3, N3=2, N4=1

Example (cont'd)

- N1=4, N2=3, N3=2, N4=1
- Revised counts $c^* = (c+1)\, N_{c+1}/N_c$
  - C*(cat) = 4
  - C*(dog) = (3+1) x 1/2 = 2
  - C*(mouse) = (2+1) x 2/3 = 2
  - C*(rabbit) = (2+1) x 2/3 = 2
  - C*(hamster) = (2+1) x 2/3 = 2
  - C*(fox) = (1+1) x 3/4 = 6/4
  - C*(turtle) = (1+1) x 3/4 = 6/4
  - C*(tiger) = (1+1) x 3/4 = 6/4
  - C*(lion) = (1+1) x 3/4 = 6/4
  - C*(elephant) = N1/N = 4/20
- Note that these counts don't necessarily add to 1, so they still need to be normalized.
  - P*(lion) = 6/4 / 20 = 6/80

Dealing with Sparse Data

- Two main techniques used
  - Backoff
  - Interpolation

Backoff

- Going back to the lower-order n-gram model if the higher-order model is sparse (e.g., frequency >= 1)
- Learning the parameters
  - From a development data set

Interpolation

- If $P'(w_i|w_{i-1},w_{i-2})$ is sparse:
  - Use $\lambda_1 P'(w_i|w_{i-1},w_{i-2})+\lambda_2 P'(w_i|w_{i-1})+\lambda_3 P'(w_i)$
- Better than backoff
- See [Chen and Goodman 1998] for more details

linear combo of tri-gram, bi-gram, uni-gram models.

NLP

6.6

Introduction to NLP

Language models (Part 3)

06.06 Language Modeling 3/3 NATURAL LANGUAGE PROCESSING

Evaluation of LM

How to evaluate the quality of a language model

- Extrinsic
  - Use in an application — translation, speech recognition
- Intrinsic
  - Cheaper — ★ Just look at the model ★
- Correlate the two for validation purposes

Perplexity

- Does the model fit the data?
  - A good model will give a high probability to a real sentence
- Perplexity
  - Average branching factor in predicting the next word
  - Lower is better (lower perplexity -> higher probability)
  - N = number of words

$$Per = \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_N)}}$$

Perplexity:
How well the model fits the data
Lower perplexity is better.

Perplexity

- Example:
  - A sentence consisting of N equiprobable words: p(wi) = 1/k

$$Per = \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_N)}}$$

  - Per = (1/k)^N^(-1/N) = k
- Perplexity is like a branching factor
- Logarithmic version:

$$Per = 2^{-(1/N)\sum \log_2 P(w_i)}$$

10 words to choose from
=) the perplexity -> 10

The Shannon Game

- Consider the Shannon game:
  - New York governor Andrew Cuomo said ...
- What is the perplexity of guessing a digit if all digits are equally likely?
  - 10
- How about a letter?
  - 26
- How about guessing A ("operator") with a probability of 1/4, B ("sales") with a probability of 1/4 and 10,000 other cases with a probability of 1/2 total (example modified from Joshua Goodman).

to predict the next word.

Predicting the next word that the customer is going to say in a call.

Perplexity Across Distributions

- What if the actual distribution is very different from the expected one?
- Example:
  - All of the 10,000 other cases are equally likely but P(A) = P(B) = 0.
- Cross-entropy = log (perplexity), measured in bits

$$H(p,q) = -\sum_x p(x) \log q(x).$$

Train with one dataset and test on another
— the distribution could be different.

Sample Values for Perplexity

- Wall Street Journal (WSJ) corpus
  - 38 M words (tokens)
  - 20 K types
- Perplexity
  - Evaluated on a separate 1.5M sample of WSJ documents
  - Unigram 962
  - Bigram 170
  - Trigram 109

20k word. words
=> unigram – less than 1000 choices.
bigram, tri-gram smaller.

Word Error Rate

- Another evaluation metric
  - Number of insertions, deletions, and substitutions
  - Normalized by sentence length
  - Same as Levenshtein Edit Distance
- Example:
  - governor Andrew Cuomo met with the mayor
  - the governor met the senator
  - 3 deletions + 1 insertion + 1 substitution = WER of 5

D D D I S

The edit distance

Issues

- Out of vocabulary words (OOV)
  - Split the training set into two parts
  - Label all words in part 2 that were not in part 1 as <UNK>
- Clustering
  - e.g., dates, monetary amounts, organizations, years

Appear in the test data that we have never seen in the training data.

cluster expressions.

Long Distance Dependencies

- This is where n-gram language models fail by definition
- Missing syntactic information
  - **The students** who participated in the game **are** tired
  - **The student** who participated in the game **is** tired
- Missing semantic information
  - **The pizza** that I had last night was **tasty**
  - **The class** that I had last night was **interesting**

missing some syntactic information.

long distance

n-gram models are going to miss this dependency.

Other Ideas in LM

Technique

- Syntactic models
  - Condition words on other words that appear in a specific syntactic relation with them
- Caching models
  - Take advantage of the fact that words appear in bursts

parse the sentence
⇒ look at the most syntactically related word before after

External Resources

- SRI-LM
  - http://www.speech.sri.com/projects/srilm/
- CMU-LM
  - http://www.speech.cs.cmu.edu/SLM/toolkit.html
- Google n-gram corpus
  - http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html
- Google book n-grams
  - http://ngrams.googlelabs.com/

Example Google n-grams

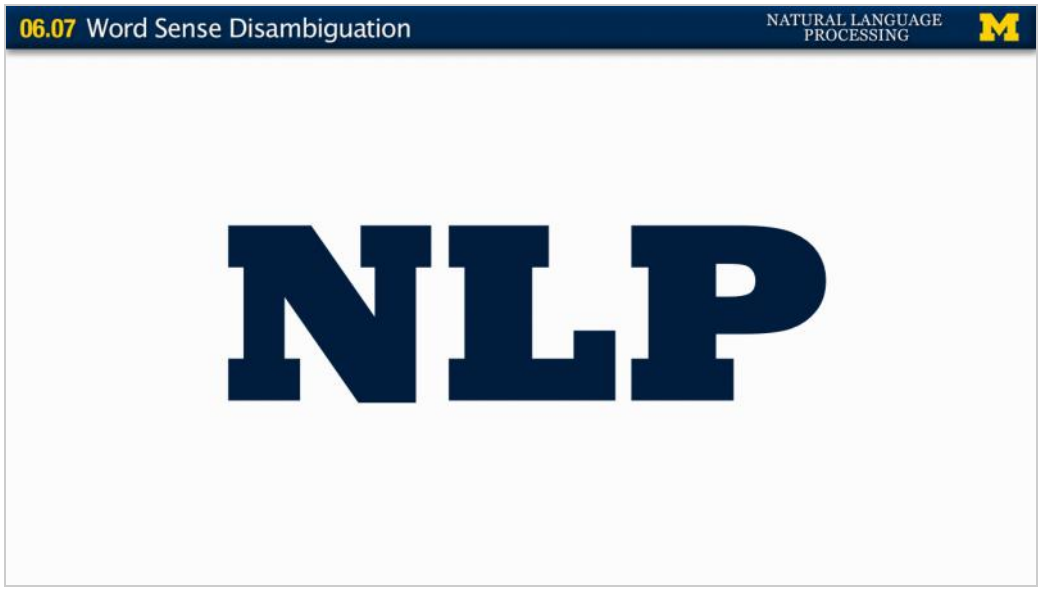
house a	302435	house hotel	139282
house after	118894	house in	3553052
house all	105970	house is	1962473
house and	3880495	house music	199346
house are	136475	house near	131889
house arrest	254629	house now	127043
house as	339590	house of	3164591
house at	694739	house on	1077835
house before	102663	house or	1172783
house built	189451	house party	162668
house but	137151	house plan	172765
house by	249118	house plans	434398
house can	133187	house price	158422
house cleaning	125206	house prices	643669
house design	120500	house rental	209614
house down	109663	house rules	108025
house fire	112325	house share	101238
house for	1635280	house so	133405
house former	112559	house that	687925
house from	249091	house the	478204
house had	154848	house to	1452996
house has	440396	house training	163056
house he	115434	house value	135820

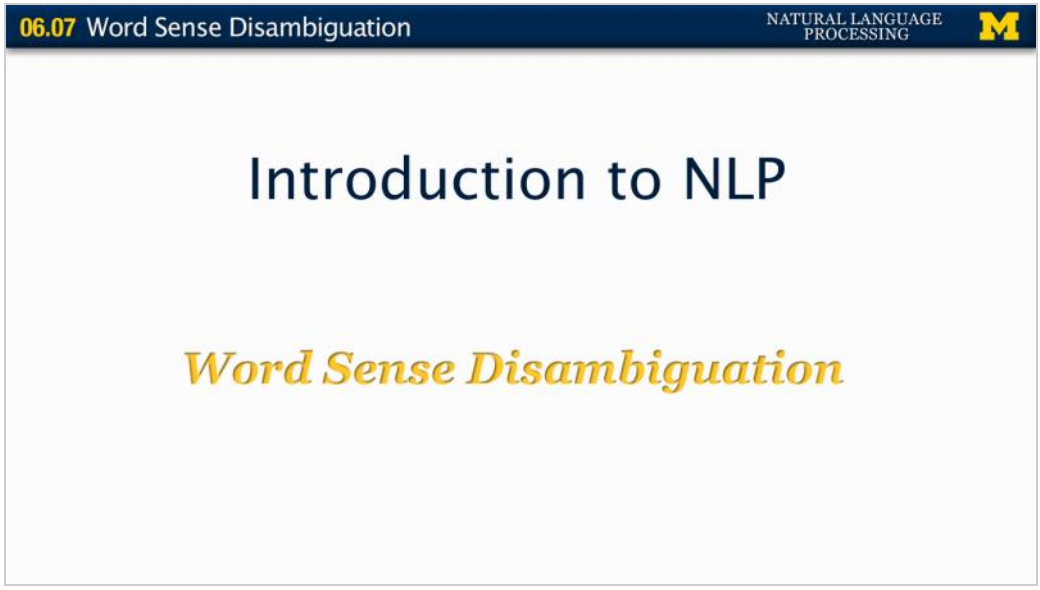
N-gram External Links

- http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html
- http://norvig.com/mayzner.html
- http://storage.googleapis.com/books/ngrams/books/datasetsv2.html
- https://books.google.com/ngrams/
- http://www.visi.com/~sgrantz/plot.html
- http://www.elsewhere.org/pomo/
- http://pdos.csail.mit.edu/scigen/
- http://www.magliery.com/Band/
- http://www.magliery.com/Country/
- http://johno.jsmf.net/knowhow/ngrams/index.php
- http://coastalweb.ca/building-sites/content-generation-with-n-grams.html
- http://www.decontextualize.com/teaching/rwet/n-grams-and-markov-chains/
- http://gregstevens.name/2012/08/16/simulating-h-p-lovecraft
- http://kingjamesprogramming.tumblr.com/

6.7

Word sense disambiguation.





Introduction

- Polysemy
  - Words have multiple senses
- Example
  - Let's have a drink in the bar
  - I have to study for the bar
  - Bring me a chocolate bar
- Homonymy
  - May I come in?
  - Let's meet again in May
- Part of speech ambiguity
  - Joe won the first round
  - Joe has a round toy

06.07 Word Sense Disambiguation NATURAL LANGUAGE PROCESSING

Senses Of The Word "Bar"

•S: (n) barroom, **bar**, saloon, ginmill, taproom (a room or establishment where alcoholic drinks are served over a counter) *"he drowned his sorrows in whiskey at the bar"*
•S: (n) **bar** (a counter where you can obtain food or drink) *"he bought a hot dog and a coke at the bar"*
•S: (n) **bar** (a rigid piece of metal or wood; usually used as a fastening or obstruction or weapon) *"there were bars in the windows to prevent escape"*
•S: (n) measure, **bar** (musical notation for a repeating pattern of musical beats) *"the orchestra omitted the last twelve bars of the song"*
•S: (n) **bar** (an obstruction (usually metal) placed at the top of a goal) *"it was an excellent kick but the ball hit the bar"*
•S: (n) prevention, **bar** (the act of preventing) *"there was no bar against leaving"; "money was allocated to study the cause and prevention of influenza"*
•S: (n) **bar** ((meteorology) a unit of pressure equal to a million dynes per square centimeter) *"unfortunately some writers have used bar for one dyne per square centimeter"*
•S: (n) **bar** (a submerged (or partly submerged) ridge in a river or along a shore) *"the boat ran aground on a submerged bar in the river"*
•S: (n) legal profession, **bar**, legal community (the body of individuals qualified to practice law in a particular jurisdiction) *"he was admitted to the bar in New Jersey"*
•S: (n) stripe, streak, **bar** (a narrow marking of a different color or texture from the background) *"a green toad with small black stripes or bars"; "may the Stars and Stripes forever wave"*
•S: (n) cake, **bar** (a block of solid substance (such as soap or wax)) *"a bar of chocolate"*
•S: (n) Browning automatic rifle, **BAR** (a portable .30 caliber automatic rifle operated by gas pressure and fed by cartridges from a magazine; used by United States troops in World War I and in World War II and in the Korean War)
•S: (n) **bar** (a horizontal rod that serves as a support for gymnasts as they perform exercises)
•S: (n) **bar** (a heating element in an electric fire) *"an electric fire with three bars"*
•S: (n) **bar** ((law) a railing that encloses the part of the courtroom where the judges and lawyers sit and the case is tried) *"spectators were not allowed past the bar"*

06.07 Word Sense Disambiguation NATURAL LANGUAGE PROCESSING

Word Sense Disambiguation

- Task
  - given a word
  - and its context
  - determine which sense it is
- Use for Machine Translation
  - e.g., translate "play" into Spanish
  - play the violin = tocar el violín
  - play tennis = jugar al tenis
- Other uses
  - Accent restoration (cote)
  - Text to speech generation (lead)
  - Spelling correction (aid/aide)
  - Capitalization restoration (Turkey)

Dictionary Method (Lesk)

- Match sentences to dictionary definitions
- Examples of plant (m-w.com):
  - $plant_1$ = a living thing that grows in the ground, usually has leaves or flowers, and needs sun and water to survive
  - $plant_2$ = a building or factory where something is made
- Examples of leaf
  - $leaf_1$ = a lateral outgrowth from a plant stem that is typically a flattened expanded variably shaped greenish organ, constitutes a unit of the foliage, and functions primarily in food manufacture by photosynthesis
  - $leaf_2$ = a part of a book or folded sheet containing a page on each side
- Find the pair of meanings that have the most overlapping definitions
  - "The *leaf* is the food making factory of green *plants*."

植物. 工厂 叶子. 书页. → the overlap in diction got the largest possibility one

06.07 Word Sense Disambiguation NATURAL LANGUAGE PROCESSING

Decision Lists (Yarowsky)

Decision Lists (Yarowsky)

- Method introduced by Yarowsky (1994)
- Two senses per word
- Ordered rules: collocation -> sense
- Formula

$$\log\left(\frac{p(\text{sense}_A \mid \text{collocation}_i)}{p(\text{sense}_B \mid \text{collocation}_i)}\right)$$

06.07 Word Sense Disambiguation NATURAL LANGUAGE PROCESSING

Decision Lists (Yarowsky)

rule

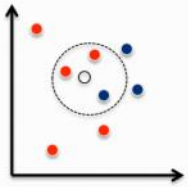
1. *fish* within window -> bass1
2. *striped bass* -> bass1
3. *guitar* within window -> bass2
4. *bass player* -> *bass2*
5. Play/V bass -> *bass2*

Classification Features

- Adjacent words (collocations)
  - e.g., chocolate bar, bar exam, bar stool, bar fight, foreign aid, presidential aide
- Position
  - e.g., plant pesticide vs. pesticide plant
- Adjacent parts of speech
- Nearby words
  - e.g., within 10 words
- Syntactic information
  - e.g., object of the verb "play"
- Topic of the text

Classification Methods

- K-nearest neighbor (memory-based)
- Using Euclidean distance
- Find the k most similar examples and return the majority class for them



Similar to other classification methods of NLP.

K-nearest neighbor.

Bootstrapping

- Start with two senses and seeds for each sense
  - e.g., plant1:leaf, plant2:factory
- Use these seeds to label the data using a supervised classifier (decision list)
- Add some of the newly labeled examples to the training data
- Repeat until no more examples can be labeled

Introduced mid-90s.

Bootstrapping -

first seed with some labeled data → get more labeled data

Bootstrapping

- Two principles:
  - one sense per collocation
  - one sense per discourse (e.g., document)

Training Data for WSD

Training data.

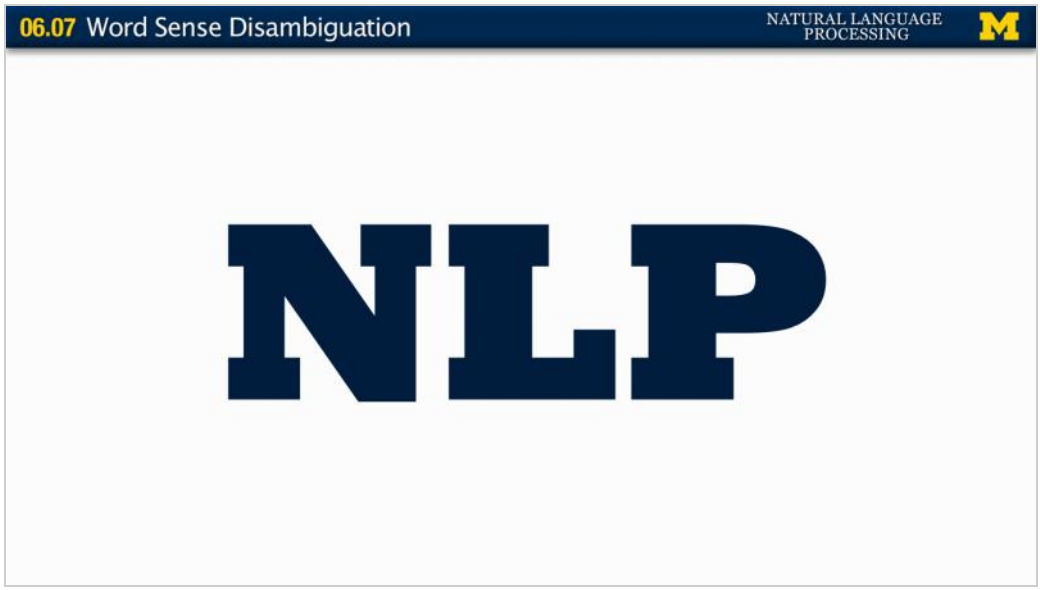
- Senseval/Semcor
  - http://www.senseval.org/senseval3
  - Lexical Sample
  - All words
  - Available for many languages
- Pseudo-words
  - E.g., banana/door
- Multilingual corpora
  - Aligned at the sentence level
  - Use the translations as an indication of sense

pseudo-word principle → take two unambiguous words => combine → train a model to guess which is the correct word. have the real data

Senseval-1 Evaluation

- Metric
  - A = number of assigned senses
  - C = number of words assigned correct senses
  - T = total number of test words
  - Precision = C/A; Recall = C/T
- Results ~77%.
  - best recall around 77P/77R
  - human lexicographer 97P/96R
  - most common sense 57P/50R (decent but depends on domain)

→ works better for text in homogeneous genre & domain.


NLP