NLP

Introduction to NLP

Language Models (1/3)

Probabilistic Language Models

- 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 be 1.

Predicting the Next Word

- Example
  - Let's meet in Times …
  - General Electric has lost some market …
- Formula
  - $P(w_n|w_1,w_2,\ldots,w_{n-1})$

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)

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")=?

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)

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 .

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

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

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
  - How many bigrams are never seen in the data?
- Notice!
  - very sparse data!

Google 1-T Corpus

- 1 trillion word tokens
- Number of tokens
  - 1,024,908,267,229
- Number of sentences
  - 95,119,665,584
- Number of unigrams
  - 13,588,391
- Number of bigrams
  - 314,843,401
- Number of trigrams
  - 977,069,902
- Number of fourgrams
  - 1,313,818,354
- Number of fivegrams
  - 1,176,470,663

https://catalog.ldc.upenn.edu/ldc2006t13

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”)

Maximum Likelihood Estimates

- 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") = ?

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 logarithms instead
  - 10^{-6} (in base 10) becomes −6
  - Use sums instead of products

Tools

- http://www.speech.cs.cmu.edu/SLM_info.html
- http://www.speech.sri.com/projects/srilm/

NLP