NLP

Introduction to NLP

Language models (3/3)

Evaluation of LM

- Extrinsic
  - Use in an application
- Intrinsic
  - Cheaper
- 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

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

Perplexity

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

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

  - Per = $((k^{-1})^N)^{(-1/N)} = k$
- Perplexity is like a branching factor
- Logarithmic version
  - the exponent is = #bits to encode each word)

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

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? Do the math.
  - 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.

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

$$\mathrm{H}(p,q) = -\sum_{x} p(x) \log q(x).$$

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

Word Error Rate

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

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

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**

Other Ideas in LM

- 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

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

NLP