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 $\rightarrow$ higher probability)
  – $N =$ number of words

$$Per = \sqrt[N]{\frac{1}{P(w_1w_2...w_N)}}$$
Perplexity

• Example:
  – A sentence consisting of N equiprobable words: \( p(w_i) = \frac{1}{k} \)

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

  – \( Per = ((k^{-1})^{\text{N}})^{(-1/N)} = k \)

• Perplexity is like a branching factor

• Logarithmic version
  – the exponent is \( = \#\text{bits to encode each word} \)

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

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

• CMU–LM
  – http://www.speech.cs.cmu.edu/SLM/toolkit.html

• Google n–gram corpus

• Google book n–grams
  – http://ngrams.googlelabs.com/
Example Google n-grams

<table>
<thead>
<tr>
<th>house a</th>
<th>302435</th>
<th>house hotel</th>
<th>139282</th>
</tr>
</thead>
<tbody>
<tr>
<td>house after</td>
<td>118894</td>
<td>house in</td>
<td>3553052</td>
</tr>
<tr>
<td>house all</td>
<td>105970</td>
<td>house is</td>
<td>1962473</td>
</tr>
<tr>
<td>house and</td>
<td>3880495</td>
<td>house music</td>
<td>199346</td>
</tr>
<tr>
<td>house are</td>
<td>136475</td>
<td>house near</td>
<td>131889</td>
</tr>
<tr>
<td>house arrest</td>
<td>254629</td>
<td>house now</td>
<td>127043</td>
</tr>
<tr>
<td>house as</td>
<td>339590</td>
<td>house of</td>
<td>3164591</td>
</tr>
<tr>
<td>house at</td>
<td>694739</td>
<td>house on</td>
<td>1077835</td>
</tr>
<tr>
<td>house before</td>
<td>102663</td>
<td>house or</td>
<td>1172783</td>
</tr>
<tr>
<td>house built</td>
<td>189451</td>
<td>house party</td>
<td>162668</td>
</tr>
<tr>
<td>house but</td>
<td>137151</td>
<td>house plan</td>
<td>172765</td>
</tr>
<tr>
<td>house by</td>
<td>249118</td>
<td>house plans</td>
<td>434398</td>
</tr>
<tr>
<td>house can</td>
<td>133187</td>
<td>house price</td>
<td>158422</td>
</tr>
<tr>
<td>house cleaning</td>
<td>125206</td>
<td>house prices</td>
<td>643669</td>
</tr>
<tr>
<td>house design</td>
<td>120500</td>
<td>house rental</td>
<td>209614</td>
</tr>
<tr>
<td>house down</td>
<td>109663</td>
<td>house rules</td>
<td>108025</td>
</tr>
<tr>
<td>house fire</td>
<td>112325</td>
<td>house share</td>
<td>101238</td>
</tr>
<tr>
<td>house for</td>
<td>1635280</td>
<td>house so</td>
<td>133405</td>
</tr>
<tr>
<td>house former</td>
<td>112559</td>
<td>house that</td>
<td>687925</td>
</tr>
<tr>
<td>house from</td>
<td>249091</td>
<td>house the</td>
<td>478204</td>
</tr>
<tr>
<td>house had</td>
<td>154848</td>
<td>house to</td>
<td>1452996</td>
</tr>
<tr>
<td>house has</td>
<td>440396</td>
<td>house training</td>
<td>163056</td>
</tr>
<tr>
<td>house he</td>
<td>115434</td>
<td>house value</td>
<td>135820</td>
</tr>
</tbody>
</table>
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.elsewhere.org/pomo/
- http://pdos.csail.mit.edu/scigen/
- http://www.magliery.com/Band/
- http://johno.jsmf.net/knowhow/ngrams/index.php
- http://coastalweb.ca/building-sites/content-generation-with-n-grams.html
- http://gregstevens.name/2012/08/16/simulating-h-p-lovecraft
- http://kingjamesprogramming.tumblr.com/
NLP