Hw5

LING 570

Fei Xia

Week 6: 10/28/09

Building trigram LM

- Step 1: collect real counts.
- Step 2: build LM
  - MLE: no smoothing
  - GT smoothing
  - …
- Step 3: calculate the perplexity
  - Use P(w3|w1,w2) alone
  - Use interpolation
  - Other options: e.g., Katz's backoff

➔ Steps 2 and 3 depend on the smoothing methods

Q1: collecting counts

- Collect real counts from the training data:
  - ngram_count.sh training_data   ngram_count_file
  - Output ngrams and real count c(w1), c(w1, w2), and c(w1, w2, w3).

- Given a sentence: John likes Mary
  - Insert BOS and EOS: <s> John likes Mary </s>
  - How many unigrams?
  - How many bigrams?
  - How many trigrams?

Output file for Q1: cnt key

895 a

….

200 the book

…

50 thank you very

…

unigrams, then bigrams, then trigrams

for each n, sort the lines by frequency

Q2: building an LM from the counts

- build_lm.sh ngram_count_file lm_file
- Store the logprob of ngrams and other parameters in the lm
  - There are actually three lm models:
    P(w3), P(w3|w2) and P(w3|w1,w2)
  - The output file is in the modified ARPA format
  - Lines for n-grams are sorted by n-gram counts

Modified ARPA format

\data
ngram 1: type=*xx* token=yy
ngram 2: type=*xx* *token=yy*
ngram 3: type=xx token=yy

\1-grams:
cnt prob logprob w # prob is P(w)
...

\2-grams:
cnt prob logprob w1 w2 # cnt is Cnt(w1,w2), prob is $P(w_2 \mid w_1)$
...
***3*-grams:**
cnt prob logprob w1 w2 w3 # prob is $P(w_3 \mid w_1, w_2)$
...
\end

Q3: calculating the perplexity

- ppl.sh lm_file λ_1 λ_2 λ_3 test_data output_file

- sum = 0;
- cnt = 0;

- for each sentence T in the test data
  for each word w_i in the sentence
  if w_i is known (aka w_i appears in the lm_file)
  $P(w3|w1,w2) = \lambda_3 P_3(w3|w1,w2) + \lambda_2 P_2(w3|w2) + \lambda_1 P_1(w3)$
  sum += log $P(w_i \mid w_{i-2}\ w_{w-1})$
  cnt ++;

- entropy = - sum / cnt
- ppl = $10^{entropy}$

➔ Q4: Calculate the perplexity with different λ's.

Output file for Q3

Sent #1: <s> Influential members of the House

1: log P(Influential | <s>) = -inf (unknown word)

2: log P(members | <s> Influential) = -4.26127986628694 (unseen ngrams)

3: log P(of | Influential members) = -0.659218767066308 (unseen ngrams)

4: log P(the | members of) = -0.673243382588536

…

37: log P(. | sick thrifts) = -2.11250099135999 (unseen ngrams)

38: log P(</s> | thrifts .) = -0.322502345739275 (unseen ngrams)

1 sentence, 37 words, 9 OOVs

logprob=-82.8860891791949 ppl=721.341645452964

…

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

sent_num=50 word_num=1175 oov_num=190

logprob=-2854.78157013778 ave_logprob=-2.75824306293506 ppl=573.116699237283

Q4: perplexity with different λ's

lambda_1	lambda_2	lambda_3	perplexity
0.05	0.15	0.8	
0.1	0.1	0.8	
0.2	0.3	0.5	
0.2	0.5	0.3	
0.2	0.7	0.1	
0.2	0.8	0	
1.0	0	0	