Word Sense Disambiguation

(SNLP tutorial 4)

Vilém Zouhar, Awantee Deshpande, Julius Steuer

TODOth, TODOth May 2021

Overview

- Task, Metrics
- Differential Privacy
- Homework

Entropy

- Amount of information / compressed size in bits
- $H(p) = E[-\log(p(V))] = -\sum p(v)\log(p(v))$
- For binomial distribution highest in the middle
- For uniform distribution: $\log(W)$
- Entropy is always non-negative
- H((W,W)) = H(W)+H(W) when statistically independent p(w1,w2) = p(w1)p(w2)
- Conditional entropy: $H(X|Y) = -\sum p(x,y)\log p(x|y)$

Kullback-Leibler Divergence

- $D(p||q) = \sum p_i \log p_i/q_i$
- Not symmetric
- Non-negative
- How many extra bits if we use bad encoding
- Cross-entropy: $-\sum p_i \log q_i$

Code

- Mapping of word to a finite string of a D-nary alphabet
- Prefix code
- $\sum D^{-l_i} \leq 1$
  - Krafts inequality
  - true for prefix codes
  - for every length distribution satisfying this, there exists a prefix code
- Expected length: $\sum l_i p(w_i)$
- Optimal length: $-\log_D p(w_i)$

Correlation Function

- $p_d(w1, w2)/(p(w1)p(w2))$

OOV words

Corpus

- Train set:

- Test set:

OOV words

Corpus

- Train set:

- Test set:

Accumulate counts

- 6 5 3 2
- 4 2 2 2 1 1

OOV words

Corpus

- Train set:

- Test set:

Accumulate counts

- 6 5 3 2
- 4 2 2 2 1 1

OOV words

- What about  and  ?
- OOV rate: $2 + 1/4 + 2 + 2 + 1 + 1 + 1 = 27\%$

Additive smoothing (add-α-smoothing)

Unigrams

- Add zero counts to frequency table

 6  5  3  2  0  0

- Increase all counts by $\alpha = 1$

 6+1  5+1  3+1  2+1  0+1  0+1

- Divide by $N = 22$

 0.32  0.27  0.18  0.13  0.05  0.05

Perplexity

- Relative frequencies on test corpus:

 0.33  0.17  0.17  0.17  0.08  0.08

Additive smoothing (add-α-smoothing)

Unigrams

- Add zero counts to frequency table

6 5 3 2 0 0

- Increase all counts by $\alpha = 1$

6+1 5+1 3+1 2+1 0+1 0+1

- Divide by $N = 22$

0.32 0.27 0.18 0.13 0.05 0.05

Perplexity

- Relative frequencies on test corpus:

0.33 0.17 0.17 0.17 0.08 0.08

- $\text{PP} = 2^{(0.33 \cdot 0.32 + 0.27 \cdot 0.17 + 0.18 \cdot 0.17 + 0.13 \cdot 0.17 + 2 \cdot (0.05 \cdot 0.08))} = 1.4$

Kneser-Ney Smoothing

TODO

- absolute discounting

Cross-Validation

TODO

Estimating LOO Parameters

TODO ??

Laplace Smoothing

- add epsilon

TODO

Linear Discounting

- linear interpolation

Good-Turing Discounting

TODO

Count Trees

- remove infrequent nodes

TODO

Privacy

TODO differential privacy

Resources

1. UdS SNLP Class, WSD: https://teaching.lsv.uni-saarland.de/snlp/
2. Classical Statistical WSD: https://www.aclweb.org/anthology/P91-1034.pdf
3. n-gram count trees: http://ssli.ee.washington.edu/WS07/notes/ngrams.pdf