CS565: Intelligent Systems and Interfaces

Language Modeling

Semester: Jan – May 2019

Ashish Anand

Associate Professor, Dept of CSE

IIT Guwahati

Announcements

- Scribe
  - Sarthak Tripathi, Abhijeet Pandey: 30th Jan Lec

- Assignment 1 posted on canvas
  - Due Date: 10th February

Recap

- In the last lecture
  - Language Modeling: Definition
  - Language Modeling: N-gram Models
  - Language Modeling: Parameter Estimation

Objective

- Discuss why STOP symbol is used ?

- Smoothing Techniques
  - Laplace or Add-1 Smoothing
  - Add-K Smoothing
  - Backoff and Interpolation methods: Generic Idea
  - Estimating parameters in Interpolation methods

MLE of N-gram models

- Unigram

$$p_{ml}(w_i) = \frac{c(w_i)}{\sum c(w_i)}$$

- Bigram

$$p_{ml}(w_i|w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

- Trigram

$$p_{ml}(w_i|w_{i-2}, w_{i-1}) = \frac{c(w_{i-2}, w_{i-1}, w_i)}{c(w_{i-2}, w_{i-1})}$$

Problem with MLE

- Works well if test corpus is very similar to training, which is not generally the case.
- Sparsity Issue
  - OOV : Can be solved by having <UNK> category
  - Words are present in corpus but relevant counts are zero
    - Underestimation of such probabilities

Smoothing Techniques

Simplest Approach: Additive Smoothing

- Add-1 Smoothing

$$p_{mle}(w_i|w_{i-2}, w_{i-1}) = \frac{c(w_{i-2}, w_{i-1}, w_i) + 1}{c(w_{i-2}, w_{i-1}) + |\mathcal{V}|}$$

- Generalized version

$$p_{mle}(w_i|w_{i-2}, w_{i-1}) = \frac{c(w_{i-2}, w_{i-1}, w_i) + \delta}{c(w_{i-2}, w_{i-1}) + \delta|\mathcal{V}|}$$

Simplest Approach: Additive Smoothing

- Few more concepts discussed in the class
  - Adjusted Count
  - Discount Ratio

What's wrong with additive smoothing

- Optional but Recommended Reading
  - Gale and Church, 1990, *Estimation procedure for language context: poor estimates are worse than none.* In *COMPSTAT*, Proceedings in Computational statistics
  - Gale and Church, 1994, *What's wrong with adding one?* Corpus-Based Research into Language.

Take the help of lower order models

- Bigram Example
  - $c(w_1, w_2) = 0 = c(w_1, w_2')$
  - $p_{add}(w_2 \mid w_1) = p_{add}(w_2' \mid w_1)$
  - Lets assume $p(w_2') < p(w_2)$
  - We should expect $p_{add}(w_2 \mid w_1) > p_{add}(w_2' \mid w_1)$

Take the help of lower order models

- Linear Interpolation Models
- Discounting Models

Linear Interpolation Model

- Bigram model $p(w_i|w_{i-1})$

$$p_{int}(w_i|w_{i-1}) = \lambda p_{mle}(w_i|w_{i-1}) + (1-\lambda)p_{mle}(w_i),$$

Where $0 \leq \lambda \leq 1$

- Trigram model

$$p_{int}(w_i|w_{i-2}, w_{i-1}) = \lambda_1 \times p_{mle}(w_i|w_{i-2}, w_{i-1}) + \lambda_2 \times p_{mle}(w_i|w_{i-1}) + \lambda_3 \times p_{mle}(w_i),$$

Linear Interpolation Model

Verify $p_{int}(w_i|w_{i-2}, w_{i-1})$ is probability distribution.

i.e., $\sum p_{int}(w_i|w_{i-2}, w_{i-1}) = 1$

Estimating λ values

- Use of validation or development or held-out data
- c'(w_1,w_2,w_3) := Number of occurrences of $w_1w_2w_3$ in the *validation data*
- Maximum likelihood estimation

$$L(\lambda_1, \lambda_2, \lambda_3) = \sum_{w_1,w_2,w_3} c'(w_1, w_2, w_3) \log p_{int}(w_3|w_1, w_2)$$

s.t. constraints on λ values.

More on Smoothing Techniques

- Optional but Recommended Reading
  - An Empirical Study of Smoothing Techniques for Language Modeling, *S Chen, and J Goodman*, 1998.

- Generalized versions
  - Interpolation Techniques
  - Discounting Methods

References

- Chapter 3 [SLP, 3rd ed. Latest draft]
- Prof. Collins Lecture Notes on Language Modeling
  - http://www.cs.columbia.edu/~mcollins/lm-spring2013.pdf