Bayesian Replacement for Good-Turing

Introduction to

MacKay (1994)"Hierarchical Dirichlet Language Model"

Daichi Mochihashi

`daichi.mochihashi@atr.jp`

ATR SLT Dept.2 Regular Meeting

May 19, 2004

Modified for Über SVM 2004

Aug 3, 2004

Overview

- N-gram smoothing is a crucial machinery in speech recognition and machine translation.
- But N-gram parameters are so numerous, and there are not much data (e.g. MAD)
  $\Downarrow$
  We can't make an exact prediction from such data.

But..

- Taking our uncertainty about the parameters into the model, we **can** make a stable prediction.
  (This is called a Bayesian Method.)
- We get a theoretically sound smoothing formula.

Introduction

By restricting ourselves to bigram for simplicity,

- Empirical (Maximum Likelihood) estimate

$$\hat{p}_{i|j} \equiv \hat{p}(w_i|w_j) = \frac{f_{i|j}}{f_j} \tag{1}$$

$(f_{i|j}, f_j$: frequency of $\langle w_j \rightarrow w_i \rangle$ and $w_j)$

  - Probability 0 for unseen words after w_j
    - e.g. $p(\mathrm{an}|\mathrm{quite}) = 0$ simply if “quite an” accidentally did not appear in the training data.
  - Some smoothing is required.

Existent smoothing

- "Adding some" method
  - Adding some count to every N-gram
  - Interpreted as an interpolation between $\hat{p}$ and uniform probability
  - Laplace smoothing, Lidstone's law, Jeffreys-Perks law, ...
- Good-Turing smoothing
  - uses "Bins of N-gram" (number of freq. 1 N-gram, ..)
  - only applicable when $f_{i|j} < \theta$.
  - shares several flaws also (next slide)

Problem of Existent smoothing

- Uniform probability to unseen words
  - $p(\mathrm{well}|\mathrm{quite}) = p(\mathrm{epistemological}|\mathrm{quite})$?
- ad hoc threshold (Good-Turing)
- frequency of context is ignored.
  - probability $0.5 = 50/100 = 2/4$?
  - the more frequent the context is, the more stable $\hat{p}$ should be (requires less smoothing)
  - But this information is discarded in the ordinary approach.

Example of Context Frequency

$$\begin{cases} \text{he} & \rightarrow \ 1000 \ \ \text{times} \\ \text{he does} & \rightarrow \ \ \ 200 \ \ \text{times} \end{cases} \quad \therefore \ p(\text{does}|\text{he}) = \frac{200}{1000} = 0.2.$$

This estimate is very reliable.

$$\begin{cases} \text{alice} & \rightarrow \ 5 \ \ \text{times} \\ \text{alice wandered} & \rightarrow \ 1 \ \ \text{time} \end{cases} \quad \therefore \ p(\text{wandered}|\text{alice}) = \frac{1}{5} = 0.2$$

$p(\text{does}|\text{he}) = p(\text{wandered}|\text{alice})$?

The latter may have been 0.3 or 0.1

$\Downarrow$

Context frequency (1000 and 5) should be considered.

Bayesian Hierarchical model

- Bigrams are governed by a probability table

$$q_{i|j} = p(w_i|w_j).$$

- But we are not confident exactly what $\mathbf{q}$ is

  $\Downarrow$

  Consider (infinite) possible $\mathbf{q}$'s, and average them.

- In fact,
  - Introducing "probability of probability table $\mathbf{q}$" and taking expectation of the prediction from each $\mathbf{q}$
  - What governs above "probability of $\mathbf{q}$" is a hyperparameter $\boldsymbol{\alpha}$ of the Dirichlet distribution.

Result of Bayesian Hierarchical model

- Resultant smoothing is a linear interpolation using empirical probability $\hat{p}_{i|j}$ and hyperparameter $\boldsymbol{\alpha}$

$$E[p(w_i|w_j)] = \frac{f_{i|j} + \alpha_i}{\sum_i (f_{i|j} + \alpha_i)} \quad (2)$$

$$= \frac{f_j}{f_j + \alpha_0} \cdot \hat{p}_{i|j} + \frac{\alpha_0}{f_j + \alpha_0} \cdot \bar{\alpha}_i \quad (3)$$

where $\alpha_0 = \sum_k \alpha_k$ and $\bar{\alpha}_i = \dfrac{\alpha_i}{\alpha_0}$

  - also depends on the frequency f_j of context w_j
  - non-uniform interpolation like back-off (α_i)
- $\bar{\alpha}_i = p(w_i)$? (unigram) $\rightarrow$ No.

Example of Bayesian model (2)

We assume $\alpha(\mathrm{does}) = 1.5,\ \alpha(\mathrm{wandered}) = 0.01,\ \alpha_0 = 10$.
Then because $f_{\mathrm{he}} = 1000$ and $f_{\mathrm{alice}} = 5$,

$$p(\mathrm{does}|\mathrm{he}) = \frac{1000}{1000+10} \cdot 0.2 + \frac{10}{1000+10} \cdot \frac{1.5}{10} = 0.1995.$$

$$p(\mathrm{wandered}|\mathrm{alice}) = \frac{5}{5+10} \cdot 0.2 + \frac{10}{5+10} \cdot \frac{0.01}{10} = 0.0673.$$

Very intuitive and different from any conventional methods that give equal probability 0.2 to both cases!

How to derive α?

- Only what remains is a hyperparameter $\boldsymbol{\alpha}$.
- Most reasonable point estimate is the $\boldsymbol{\alpha}$ which maximizes the probability of observed counts $F = \{f_{i|j}\}$ (called "evidence" in Bayesian statistics)

$$p(F|\boldsymbol{\alpha}) = \int p(F|\mathbf{q})p(\mathbf{q}|\boldsymbol{\alpha})d\mathbf{q}$$

$$= \prod_{j=1}^{L} \int_0^1 \cdots \int_0^1 \prod_{i=1}^{L} q_i^{f_{i|j}} \cdot \frac{\Gamma(\alpha_0)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{L} q_i^{\alpha_i - 1} dq_1 \cdots dq_L$$

$$= \prod_{j=1}^{L} \left[\frac{\Gamma(\alpha_0)}{\prod_i \Gamma(\alpha_i)} \cdot \frac{\prod_i \Gamma(f_{i|j} + \alpha_i)}{\Gamma(f_j + \alpha_0)} \right] \quad (4)$$

How to derive α? (2)

$$p(F|\boldsymbol{\alpha}) = \prod_{j=1}^{L} \left[\frac{\Gamma(\alpha_0)}{\prod_i \Gamma(\alpha_i)} \cdot \frac{\prod_i \Gamma(f_{i|j} + \alpha_i)}{\Gamma(f_j + \alpha_0)} \right] \tag{5}$$

- This evidence (likelihood) is convex in α, and has a global maximum
- Maximum of α can be obtained by an iterative optimization (MacKay 1994, Minka 2003)
  - 77 lines of MATLAB code last week
  - Taking a few hours to calculate (for small data).

Minka's Exact Method

- Minka (2003) "Estimating a Dirichlet distribution"

$$\alpha_i^{(t+1)} = \alpha_i^{(t)} \cdot \frac{\sum_j \Psi(f_{i|j} + \alpha_j) - \Psi(\alpha_j)}{\sum_j \Psi(f_j + \sum_k \alpha_k) - \Psi(\sum_k \alpha_k)}. \qquad (6)$$

where $\Psi(x) = \frac{d}{dx} \log \Gamma(x)$

- For bigrams, it takes about 30 minutes on P4 2GHz (dependent on data)
- MATLAB code available on request.

MacKay's Approximation

- MacKay (1994) approximates $\Psi(x)$ by expansion:

$$K(\alpha) = \sum_{j=1}^{L} \log \frac{f_j + \alpha}{\alpha} + \frac{1}{2} \sum_{j=1}^{L} \frac{f_j}{\alpha(f_j + \alpha)} \tag{7}$$

$$V(i) = \text{number of contexts before word } i$$

$$G(i), H(i) = \text{sufficient statistics from the n-gram table}$$

Then, (no proof is given!)

$$\alpha_i' = 2V(i) / \left[K(\alpha_i) - G(i) + \sqrt{(K(\alpha_i) - G(i))^2 + 4H(i)V(i)} \right] \tag{8}$$

- Consistent to the exact answer (while difference of performance needs to be examined.)

Is it perfect?

- Yes, almost perfect.
- But, in general history h, the formula is:

$$E[p(w_i|h)] = \frac{f_h}{f_h + \alpha_0} \cdot \hat{p}_{i|h} + \frac{\alpha_0}{f_h + \alpha_0} \cdot \bar{\alpha}_i \tag{9}$$

- It uses a MLE (no smoothing) for history frequency f_h !
- We must estimate f_h *recursively* also by a Bayesian method. (current work)
  - Due to the point estimate of hyperparameter and the assumption of uniform hyperprior.

Gamma function

- Gamma function $\Gamma(x)$ is a continuous analogue of the factorial
- $\Gamma(x) = (x-1)!$ if x is an integer
- $\Gamma(x)$ is defined by: $\Gamma(x) = \int_0^\infty \exp(-\theta)\theta^{x-1} d\theta$.