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)

  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

\[ \hat{p}_{i|j} \equiv \hat{p}(w_i|w_j) = \frac{f_{i|j}}{f_j} \]  

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

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

- 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

\[
\begin{align*}
\{ & \text{he} \rightarrow 1000 \text{ times} \quad \therefore p(\text{does}|\text{he}) = \frac{200}{1000} = 0.2. \\
& \text{he does} \rightarrow 200 \text{ times} \\
\{ & \text{alice} \rightarrow 5 \text{ times} \quad \therefore p(\text{wandered}|\text{alice}) = \frac{1}{5} = 0.2 \\
& \text{alice wandered} \rightarrow 1 \text{ time} \\
\end{align*}
\]

This estimate is very reliable.

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

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 \( q \) is

\[ \Downarrow \]

Consider (infinite) possible \( q \)'s, and average them.

In fact,

- Introducing “probability of probability table \( q \)” and taking expectation of the prediction from each \( q \)
- What governs above “probability of \( q \)” is a hyperparameter \( \alpha \) of the Dirichlet distribution.
Result of Bayesian Hierarchical model

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

$$E[p(w_i|w_j)] = \frac{f_{i|j} + \alpha_i}{\sum_i (f_{i|j} + \alpha_i)}$$  \hspace{1cm} (2)

$$= \frac{f_j}{f_j + \alpha_0} \cdot \hat{p}_{i|j} + \frac{\alpha_0}{f_j + \alpha_0} \cdot \bar{\alpha}_i$$  \hspace{1cm} (3)

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

- also depends on the frequency $f_j$ of context $w_j$
- non-uniform interpolation like back-off ($\alpha_i$)

$\bar{\alpha}_i = p(w_i)$? (unigram) $\rightarrow$ No.
Example of Bayesian model (2)

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

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

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

Very intuitive and different from any conventional methods that give equal probability 0.2 to both cases!
How to derive $\alpha$?

Only what remains is a hyperparameter $\alpha$.

Most reasonable point estimate is the $\alpha$ which maximizes the probability of observed counts $F = \{f_{i|j}\}$ (called “evidence” in Bayesian statistics)

$$p(F|\alpha) = \int p(F|q)p(q|\alpha)dq$$

$$= \prod_{j=1}^{L} \int_{0}^{1} \cdots \int_{0}^{1} \prod_{i=1}^{L} q_{i}^{f_{i|j}} \cdot \frac{\Gamma(\alpha_{0})}{\prod_{i} \Gamma(\alpha_{i})} \prod_{i=1}^{L} q_{i}^{q_{i}^{\alpha_{i}-1}d_{q_{1}} \cdots d_{q_{L}}$$

$$= \prod_{j=1}^{L} \left[ \frac{\Gamma(\alpha_{0})}{\prod_{i} \Gamma(\alpha_{i})} \cdot \frac{\prod_{i} \Gamma(f_{i|j} + \alpha_{i})}{\Gamma(f_{j} + \alpha_{0})} \right]$$

(4)
How to derive $\alpha$? (2)

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

- This evidence (likelihood) is convex in $\alpha$, and has a global maximum
- Maximum of $\alpha$ 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


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

(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:

$$K(\alpha) = \sum_{j=1}^{L} \log \frac{f_j + \alpha}{\alpha} + \frac{1}{2} \sum_{j=1}^{L} \frac{f_j}{\alpha(f_j + \alpha)}$$  \hspace{1cm} (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!)

$$\alpha'_i = \frac{2V(i)}{K(\alpha_i) - G(i) + \sqrt{(K(\alpha_i) - G(i))^2 + 4H(i)V(i)}}$$  \hspace{1cm} (8)

Consistent to the exact answer (while difference of performance needs to be examined.)
Yes, almost perfect.

But, in general history $h$, the formula is:

$$E[p(w_i|h)] = \frac{f_h}{f_h + \alpha_0} \cdot \hat{p}_{i|h} + \frac{\alpha_0}{f_h + \alpha_0} \cdot \bar{\alpha}_i$$  \hspace{1cm} (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: $\Gamma(x) = \int_0^\infty \exp(-\theta)\theta^{x-1} d\theta$. 