N-gram Frequency Discounts

Paul Glezen

March 4, 2017

This is just a short note to straighten out some notational confusion I encountered while grappling with the Good-Turing estimate for frequency discounts employed within the Katz backoff model. Let's say we have a collection of n-grams that we have counted and aggregated into n-grams types. Each *type* consists of different occurrences of the same sequence of n tokens. I defined the following notation.

- C – The count of all n-gram occurrences
- N – The number of distinct n-grams types
- N_c – The number of distinct n-gram types that occur with count c
- F – The largest count for an n-gram type.

So, for example, the expression $N_4 = 7$ means that there are seven n-gram types that have four n-gram occurrences. An important quantity is N_1, the number of n-gram types that occur only once.

We then have the following relations for n-gram counts and frequency counts. For the count of all the distinct types

$$N = \sum_{c=1}^{F} N_c$$

For the count of all occurrences

$$C = \sum_{c=1}^{F} cN_c$$

The probability of encountering another occurrence of an n-gram we've seen before is done in the usual way. Let t_i designate the ith type and c_i the count of the ith type. Then

$$P(\text{type} = t_i) = \frac{c_i}{C}$$

where $C = \sum_{j=1}^{N} c_j$. This is the count for a type divided by the count across all types.

We want to address the problem of predicting the probability of encountering an n-gram we haven't encountered. We estimate this to be the number of n-gram types we've seen only once divided by the total count of n-grams.

$$q_0 = \frac{N_1}{C}$$

We wish to "set aside" this probability, which means we need to take it from somewhere else. Let's take an equal *glob* g from each n-gram type count c_i so that the total probability still equals 1.

$$\begin{aligned}
1 &= \frac{N_1}{C} + \sum_{i=1}^{N} \frac{c_i - g}{C} \\
&= \frac{N_1}{C} + \sum_{i=1}^{N} \frac{c_i}{C} - \sum_{i=1}^{N} \frac{g}{C} \\
&= \frac{N_1}{C} + \frac{1}{C}\sum_{i=1}^{N} c_i - \frac{g}{C}\sum_{i=1}^{N} 1 \\
&= \frac{N_1}{C} + \frac{C}{C} - \frac{g}{C}N \\
1 &= \frac{N_1}{C} + 1 - g\frac{N}{C}
\end{aligned}$$

Subtract 1 from both sides and solve for g.

$$g = \frac{N_1}{N}$$

Thus the size of the glob g to subtract from each n-gram type count is N_1/N where N is the number of n-gram types and N_1 is the number of n-gram types that occur only once.

Acknowledgements

- *Katz's back-off model*. Wikipedia. https://en.wikipedia.org/wiki/Katz's_back-off_model

- *Discounting Methods*. Columbia University Course: Natural Language Processing. https://www.youtube.com/watch?v=hsHw9F3UuAQ. My derivation was inspired by the example described in this video.

- Michael Szczepaniak pointed me to the above video.