University of Malta
BSc IT (Hons)Year IV

CSA4050: Advanced Topics Natural Language Processing

Lecture Statistics III

Statistical Approaches to NLP

- Witten-Bell Discounting
- Unigrams
- Bigrams

Dept Computer Science and AI 2003/04
Lecturer: Michael Rosner

Add-One Smoothing

Add-one smoothing is not a very good method because

- Too much probability mass is assigned to previously unseen bigrams.
- Too little probability is assigned to those bigrams having non-zero counts.
- Add-one is worse at predicting zero-count bigrams than other methods.
- Variances of the counts produced by the add-one method are actually worse than than those for the unsmoothed method.

Witten-Bell (1991) Discounting

- Assume that a zero-frequency event is one that has not happened *yet*
- When it does happen, it will be the first time it happens.
- So the probability of seeing a zero-frequency N-gram can be modelled by the probability of seeing an N-gram for the first time.
- **Key Idea**: Use the count of things you've seen *just once* to help estimate the count of things you've never seen.
- What are the things we have seen just once?

Probability of N-Grams Seen Once

- The count of N-grams seen just once is the same as the number of N-gram *types* that have been seen in the corpus.
- We therefore estimate the the *total* probability mass of *all* the zero N-grams as

$$\frac{T}{N+T}$$

  - T is the number of *observed* types.
  - N is the number of *observed* tokens.
- Divide by Z, the number of unseen N-grams, to get the probability per unseen N-gram

$$\frac{T}{Z(N+T)}$$

Discounting

- Multiplying this probability by N yields c_i^*

$$c_i^* = N\frac{T}{Z(N+T)}$$

the adjusted count for zero-count N-grams.

- Now, the total probability mass assigned to unseen N-grams $\frac{T}{N+T}$, came by *discounting* the probability of the *seen* N-grams. Clearly, the remaining probability mass is

$$1 - \frac{T}{N+T} = \frac{N}{N+T}$$

- This multiplier can be used for obtaining
  - discounted probabilities $p_i^* = (\frac{N}{N+T})p_i$,
  - smoothed counts $c_i^* = (\frac{N}{N+T})c_i$

  for the seen N-grams.

Summary for Unigrams

$$p_i^* = \begin{cases} \frac{T}{Z(N+T)} & \text{if } c_i = 0 \\ p_i \frac{N}{N+T} & \text{if } c_i > 0 \end{cases}$$

We now extend the analysis to bigrams.

Witten-Bell for Bigrams

- Generalise

  seeing a unigram for the first time ⇒ seeing a bigram *with a given first member* for the first time.

- Probability of an unseen $w_x w_k$ is estimated using the probability of bigrams $w_x w_i$ *actually* seen for the first time, i.e the number of actual $w_x w_i$ types.

- Not all conditioning events are equally informative, e.g. "Coca ..." vs. "the ..."

  There are very few "Coca x" types compared to "the x" types.

- Therefore we should give lower estimate for unseen bigrams to contexts where very few distinct word types follow a conditioning event.

Witten-Bell for Bigrams
Unseen Bigrams 1

For the collection of unseen bigrams $w_x w_i$, the total probability mass to be assigned is:

$$\sum_{i:c(w_x w_i)=0} = p(w_i \mid w_x) = \frac{T(w_x)}{N(w_x) + T(w_x)}$$

where

- N(w): number of observed bigram tokens beginning with w
- T(w): number of observed bigram types beginning with w

Witten-Bell for Bigrams 2
Unseen Bigrams 2

To obtain adjusted probability per unseen bigram:

Divide by $Z(W_x)$ – the number of unseen bigrams beginning with w_x.

$$p^*(w_i \mid w_x) = \frac{T(w_x)}{Z(w_x) \quad (N(w_x) + T(w_x))}$$

Since we know the vocabulary size V (total number of word types), we know the number of theoretically possible bigrams beginning with w_x.

We also know $T(w_x)$ the number of *observed* bigrams beginning with x.

From these we can infer

$$Z(W_x) = V - T(w_x)$$

Witten-Bell for Bigrams 3

The probability mass transferred to unseen bigrams must now be discounted from the observed ones, whose adjusted probability is:

$$p^*(w_i \mid w_x) = \frac{c(w_x w_i)}{c(w_x) + T(w_x)}$$

where $T(w_x)$ is the number of bigrams tokens beginning with w_x

Summary for Bigrams

$$p^*(w_i \mid w_x) =$$

$$\begin{cases} \frac{T(w_x)}{Z(w_x) \quad (N(w_x)+T(w_x))} & \text{if } c(w_x w_i) = 0 \\ \frac{c(w_x w_i)}{c(w_x)+T(w_x)} & \text{if } c(w_x w_i) > 0 \end{cases}$$