

# NLP - Fall 2016 - Midterm Exam

This material is ©Anoop Sarkar 2016.

Only students registered for this course are allowed to download this material.

Use of this material for “tutoring” is prohibited.

## (1) Probability Models and Smoothing:

- a. (4pts) You are given the following training data for the prepositional phrase (PP) attachment task.

v	n1	p	n2	Attachment
join	board	as	director	V
is	chairman	of	N.V.	N
using	crocidolite	in	filters	V
⋮	⋮	⋮	⋮	⋮

In order to resolve PP attachment ambiguity we can train a probability model:  $P(A = N | v, n1, p, n2)$  which predicts the attachment  $A$  as  $N$  if  $P > 0.5$  and  $V$  otherwise. Since we are unlikely to see the same four words  $v, n1, p, n2$  in novel unseen data, in order for this probability model to be useful we need to take care of zero counts.

Provide a Jelinek-Mercer *interpolation* smoothing model  $\hat{P}(A = N | v, n1, p, n2)$  for this PP attachment probability model. Assume that our training data is large enough to contain all the prepositions we might observe in unseen data. You cannot assume that all the verbs and nouns in the unseen data were seen in training.

In the interpolation model if you use any new variables then provide the constraints that the variables must obey such that  $\hat{P}$  continues to be a valid probability.

*Answer:*

$$\begin{aligned}\hat{P}(A = N | v, n1, p, n2) = & \lambda_1 P(A = N | v, n1, p, n2) \\ & + \lambda_2 P(A = N | v, n1, p) \\ & + \lambda_3 P(A = N | n1, p) \\ & + \lambda_4 P(A = N | v, p) \\ & + \lambda_5 P(A = N | p)\end{aligned}$$

To be a well-formed interpolation model,  $\sum_i \lambda_i = 1$ . There are many other solutions such as for instance, the 4-gram model can be recursively interpolated with the 3-gram model and so on just like the JM model for n-grams.

- b. (2pts) You are given a text of words:  $w_1, \dots, w_N$  and you proceed to estimate bigram probabilities with the maximum likelihood estimate using the bigram frequencies  $c(w_i, w_{i-1})$  and unigram frequencies  $c(w_i)$ :

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

The frequency for all individual words  $w_i$  are non-zero, i.e.  $c(w_i) > 0$  for all  $w_i$  but many bigrams  $w_i, w_{i-1}$  have zero counts. To avoid zeroes in the numerator of the above equation, you decide to add a small factor  $\delta$  to each count in the following manner:

$$\hat{P}(w_i | w_{i-1}) = \frac{\delta + c(w_i, w_{i-1})}{c(w_{i-1})}$$

Where  $0 < \delta < 1$ . Is this equation correct? If not, what is the basic condition on  $P(w_i | w_{i-1})$  violated in this formula? If incorrect, provide a correction for this formula.

*Answer:* The following condition is violated:

$$1 = \sum_{w_i} \hat{P}(w_i | w_{i-1})$$

Assuming a vocabulary size of  $V$  the number of  $\delta$  terms added across all the numerators will be  $\delta V$  so to make it sum to 1:

$$\hat{P}(w_i | w_{i-1}) = \frac{\delta + c(w_i, w_{i-1})}{\delta V + c(w_i)}$$

c. (4pts) For bigram probabilities, backoff smoothing is defined as follows:

$$P_{bo}(w_i | w_{i-1}) = \begin{cases} \frac{c^*(w_{i-1}, w_i)}{c(w_{i-1})} & \text{if } c(w_{i-1}, w_i) > 0 \\ \alpha(w_{i-1}) P_{bo}(w_i) & \text{otherwise} \end{cases}$$

where  $\alpha(w_{i-1})$  is chosen to make sure that  $P_{bo}(w_i | w_{i-1})$  is a proper probability

$$\alpha(w_{i-1}) = 1 - \sum_{w_i} \frac{c^*(w_{i-1}, w_i)}{c(w_{i-1})}$$

Provide the definition for the new count  $c^*(w_{i-1}, w_i)$  using the absolute discounting method. Assume that  $1 = \sum_{w_i} P_{bo}(w_i)$ .

*Answer:*

**Good Turing**

$$c^*(w_{i-1}, w_i) = (c(w_{i-1}, w_i) + 1) \times \frac{n_{c(w_{i-1}, w_i)} + 1}{n_{c(w_{i-1}, w_i)}}$$

$$c^*(w_i) = (c(w_i) + 1) \times \frac{n_{c(w_i)} + 1}{n_{c(w_i)}}$$

where  $n_{c(w_{i-1}, w_i)}$  and  $n_{c(w_i)}$  stands for the number of different  $w_{i-1}, w_i$  or  $w_i$  types observed for count  $c(w_{i-1}, w_i)$  and  $c(w_i)$  respectively.

**Absolute Discounting**

$$c^*(w_{i-1}, w_i) = c(w_{i-1}, w_i) - D$$

where  $D$  is set to some value less than one using held out set.

## (2) Word Alignment and Machine Translation:

IBM Model 1 provides the probability of an alignment  $\mathbf{a}$  and source sentence  $\mathbf{f}$  given a target sentence  $\mathbf{e}$  using the parameters  $t$  follows:

$$\Pr(\mathbf{f}, \mathbf{a} | \mathbf{e}) = \prod_{i=1}^I t(f_i | e_{a_i})$$

a. (5pts) In order to compute the most likely alignment  $\mathbf{a}$  we need to compute  $P(\mathbf{a} | \mathbf{f}, \mathbf{e})$ . Derive the following formula and provide the value of  $Z$  using parameters  $t$ :

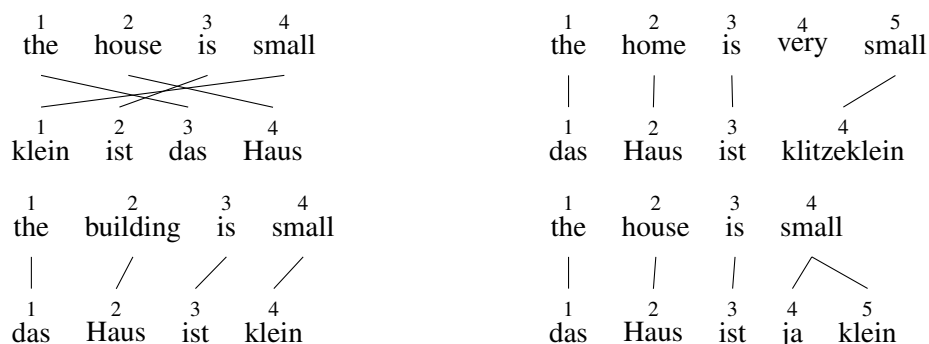
$$\log \Pr(\mathbf{a} | \mathbf{f}, \mathbf{e}) = \sum_{i=1}^I \log t(f_i | e_{a_i}) - \log Z$$

Answer:

$$\log \Pr(\mathbf{a} \mid \mathbf{f}, \mathbf{e}) = \log \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) - \log Z$$

$$\begin{aligned} \Pr(\mathbf{a} \mid \mathbf{f}, \mathbf{e}) &= \frac{\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})}{\Pr(\mathbf{f} \mid \mathbf{e})} \\ &= \frac{\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})}{\sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})} \\ &= \frac{\prod_{i=1}^I t(f_i \mid e_{a_i})}{\underbrace{\sum_{\mathbf{a}} \prod_{i=1}^I t(f_i \mid e_{a_i})}_Z} \end{aligned}$$

b. (5pts) Consider the following sentence aligned and word aligned parallel corpus.



Provide the values of the following parameters:

i  $t(\text{klein} \mid \text{small})$

Answer:  $\frac{3}{5}$

ii  $t(\text{klitzeklein} \mid \text{small})$

Answer:  $\frac{1}{5}$

iii  $t(\text{ja} \mid \text{small})$

Answer:  $\frac{1}{5}$

iv  $t(\text{ja} \mid \text{very})$

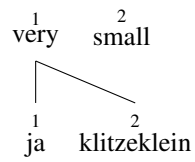
Answer: 0

v  $t(\text{klitzeklein} \mid \text{very})$

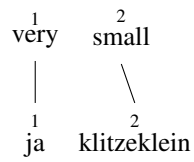
Answer: 0

- c. (5pts) Compute  $\Pr(\mathbf{ja\ klitzeklein} \mid \mathbf{very\ small})$  using IBM Model 1 without NULLs. Show all the steps in the computation and write your answer as a fraction  $\frac{x}{y}$ .

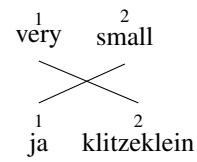
*Answer:* First compute  $\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$  for each alignment  $\mathbf{a}$ :



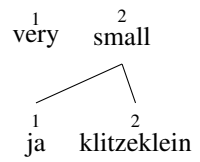
$$0 \times 0 = 0$$



$$0 \times \frac{1}{5} = 0$$



$$\frac{1}{5} \times 0 = 0$$



$$\frac{1}{5} \times \frac{1}{5} = \frac{1}{25}$$

Then sum over all alignments to get  $\Pr(\mathbf{f} \mid \mathbf{e})$ :

$$0 + 0 + 0 + \frac{1}{25} = \frac{1}{25}$$