

# NLP - Fall 2019 - Sample Midterm Exam

This material is ©Anoop Sarkar 2019.

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

Use of this material for “tutoring” is prohibited.

## (1) Language Models

Ms. Malaprop would like to build a spelling corrector focused on the particular problem of *there* vs *their*. The idea is to build a model that takes a sentence as input, for example:

1. He saw their football in the park
2. He saw their was a football in the park

For each instance of *their* or *there* Ms. Malaprop wants to predict whether the true spelling should be *their* or *there*. So for sentence (1) the model should predict *their*, and for sentence (2) the model should predict *there*. Note that for the second example the model would correct the spelling mistake in the sentence. Ms. Malaprop recently took some NLP classes so she wants to use a language model for this task. Given a language model  $p(w_1, \dots, w_n)$ , return the spelling that gives the highest probability under the language model. So for example for the second sentence we would implement the rule: replace *there* with *their* and vice versa and compare the language model scores:

```
If    p(He saw there was a football in the park) >
      p(He saw their was a football in the park)
Then  return(there)
Else  return(their)
```

Ms. Malaprop decides to use an unigram model:  $p(w_1, \dots, w_n) = \prod_{i=1}^n q(w_i)$  where  $q(w_i) = \frac{\text{Count}(w_i)}{N}$  and  $N = \sum_w \text{Count}(w)$ .  $\text{Count}(\cdot)$  returns the number of times a word was seen in the corpus and  $N$  is the sum of counts for all words in the corpus. Assume  $N = 10,000$  and  $\text{Count}(\text{there}) = 110$  and  $\text{Count}(\text{their}) = 50$ . Also assume that for every word  $w$  in the vocabulary  $\text{Count}(w) > 0$ .

- a. What does the Ms. Malaprop rule return for *He saw their was a football in the park*?

Answer: *there*

- max mark: 1 (all or nothing)

- b. Is the Ms. Malaprop rule a good solution to the *their* versus *there* problem? Say yes or no and give a short and precise one sentence justification for your answer.

Answer: No. Gives the wrong answer for *He saw their football in the park*.

- max mark: 4
- yes / no: 1 mark
- explanation makes sense: 3 marks

## (2) Smoothing:

Let  $c_{i+k}^i$  represent a sequence of characters  $c_i, c_{i+1}, \dots, c_{i+k}$ . You are given a 4-gram character model:  $P(c_i | c_{i-1}^{i-3})$ . Assume that all the characters have been observed at least once in the training data such that  $P(c_i)$  is never zero in unseen data.

- a. Consider a backoff smoothing model  $\hat{P}$  which deals with events that have been observed zero times in the training data:

$$\hat{P}(c_i | c_{i-1}^{i-3}) = \begin{cases} P(c_i | c_{i-1}^{i-3}) & \text{if } f(c_i^{i-3}) > 0 \\ P(c_i | c_{i-1}^{i-2}) & \text{if } f(c_i^{i-3}) = 0 \text{ and } f(c_i^{i-2}) > 0 \\ P(c_i | c_{i-1}) & \text{if } f(c_i^{i-2}) = 0 \text{ and } f(c_i^{i-1}) > 0 \\ P(c_i) & \text{otherwise} \end{cases}$$

where  $f(c_{i+k}^i)$  is the number of times the n-gram  $c_{i+k}^i$  was observed in the training data. What condition that holds in the original model  $P(c_i | c_{i-1}^{i-3})$  is violated by  $\hat{P}(c_i | c_{i-1}^{i-3})$ ?

Answer:

$$\sum_{c_i} P(c_i | c_{i-1}^{i-3}) = 1$$

- b. The recursive definition of Jelinek-Mercer style interpolation smoothing is:

$$\hat{P}_4(c_i | c_{i-1}^{i-3}) = \lambda_3 \cdot P(c_i | c_{i-1}^{i-3}) + (1 - \lambda_3) \cdot \hat{P}_3(c_i | c_{i-1}^{i-2})$$

Note the difference between  $P(\cdot)$  and  $\hat{P}(\cdot)$  where  $P(\cdot)$  is computed using the character frequencies in the training corpus. Based on the definition of  $\hat{P}_4(\cdot)$  provide the definitions for  $\hat{P}_3(\cdot)$ ,  $\hat{P}_2(\cdot)$ , and  $\hat{P}_1(\cdot)$  using the interpolation parameters  $\lambda_2, \lambda_1$ .

Answer:

$$\begin{aligned} \hat{P}_3(c_i | c_{i-1}^{i-2}) &= \lambda_2 \cdot P(c_i | c_{i-1}^{i-2}) + (1 - \lambda_2) \cdot \hat{P}_2(c_i | c_{i-1}) \\ \hat{P}_2(c_i | c_{i-1}) &= \lambda_1 \cdot P(c_i | c_{i-1}) + (1 - \lambda_1) \cdot \hat{P}_1(c_i) \\ \hat{P}_1(c_i) &= P(c_i) \end{aligned}$$

- c. State the condition on values assigned to  $\lambda_3, \lambda_2, \lambda_1$  in Question 2b in order to ensure that  $\hat{P}_4(c_i | c_{i-1}^{i-3})$  is a well-defined probability model.

Answer:

$$\text{where } 0 \leq \lambda_i \leq 1, i = 1, 2, 3$$

- d. Consider the bigram character model  $P(c_i | c_{i-1})$  where

$$P(c_0, \dots, c_n) = P(c_0) \times \prod_{i=1,2,\dots,n} P(c_i | c_{i-1}) \quad (1)$$

Based on this model, for the English word *booking* the probability would be computed as:

$$P(\text{booking}) = P(b) \times P(o | b) \times P(o | o) \times P(k | o) \times P(i | k) \times P(n | i) \times P(g | n)$$

Let us assume we want to generate the suffix *ing* after the stem *book* using a separate probability:

$$P(\text{ing}) = P(i) \times P(n | i) \times P(g | n)$$

In Semitic languages, like Arabic and Hebrew, the process of inflection works a bit differently. In Arabic, for a word like *kitab* the stem would be *k-t-b* where the place-holders ' ' for inflection characters have been added for convenience. We will assume that each word is made up of a sequence of consonant-vowel sequences CVCVCV... and the vowels always form the inflection.

Provide the definition of an  $n$ -gram model that will compute the probability for the word *kitab* and *k-t-b* as follows:

$$P(kitab) = P(k) \times P(t | k) \times P(b | t) \times P(i) \times P(a | i)$$

$$P(k-t-b) = P(k) \times P(t | k) \times P(b | t) \times P(-) \times P(- | -)$$

Write down the equation for this  $n$ -gram model in the same mathematical notation as equation (1). You can assume that the input will be longer than some minimum length.

*Answer:*

$$P(c_0, \dots, c_n) = P(c_0) \times \prod_{i=2,4,\dots,n} P(c_i | c_{i-2}) \times \left( P(c_1) \times \prod_{i=3,5,\dots,n-1} P(c_i | c_{i-2}) \right)$$