# Quiz 2: Example Questions

Anoop Sarkar – anoop@cs.sfu.ca

## 1 Topics

The topics for Quiz 2 are:

1. Zipf's Law
2. Relation between edit-distance and transducers
3. Noisy-channel model
   - Spelling correction
   - Part of speech tagging
4. Language models
   - $n$-gram models, Markov chains
   - Smoothing $n$-grams
5. Hidden Markov Models
   - State sequence vs. observation sequence (what is hidden?)
   - Viterbi algorithm (finding best state sequence)

## 2 Example Questions

(1) Zipf's formula $f \propto \frac{1}{r}$ and Mandelbrot's formula $f = P(r + \rho)^{-B}$ relate the sorted rank $r$ of each word in a corpus with the word's frequency $f$. Under what condition are they equivalent.

(2) Given the alphabet $\{a, b, c, \ldots, z, ␣\}$, where ␣ is the space character, provide a transducer that implements the traditional rule of spelling correction — "*i before e except after c*". Use it to correct the inputs *'yeild'* and *'reciept'* (provide the state sequence in your transducer, the input word and the output word). Check if your transducer works correctly on the input word *'either'*. You can use Perl character classes to generalize over groups of letters from the alphabet, e.g. `[a-z,␣]` stands for any letter in the alphabet and `[^a]` stands for any letter in the alphabet except `a`.

(3) Let $c^i_{i+k}$ represent a sequence of characters $c_i, c_{i+1}, \ldots, c_{i+k}$. Assume that you're given a 4-gram character model: $P(c_i \mid c^{i-3}_{i-1})$. Note that $\sum_{c_i} P(c_i \mid c^{i-3}_{i-1}) = 1$. 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:

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

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 is violated by $\hat{P}$?

b. Consider a Jelinek-Mercer style interpolation smoothing model $\hat{P}$:

$$\hat{P}(c_i \mid c_{i-1}^{i-3}) = \lambda_1 \cdot P(c_i \mid c_{i-1}^{i-3}) + \lambda_2 \cdot P(c_i \mid c_{i-1}^{i-2}) + \lambda_3 \cdot P(c_i \mid c_{i-1}) + \lambda_4 \cdot P(c_i)$$

State the condition on values assigned to $\lambda_1, \ldots, \lambda_4$ for $\hat{P}$ to be a well-defined probability model.

c. Assume you are given some additional training data (separate from your original training data). Let's say this data is your *held-out* data called $T$. $T$ will contain ngrams that were unseen in our original training data, and we can exploit this fact to compute values for $\lambda_i$ in an interpolation smoothing model.

Let $W_T$ be the length of $T$ in number of character tokens. For each *token* $t_i$ in T, where $1 \leq i \leq W_T$, let $g_1(t_i) = 1$ iff the 4-gram probability $P(t_i \mid t_{i-1}^{i-3})$ had a non-zero value (that is, whenever $f(t_i^{i-3}) > 0$). Similarly, let $g_2(t_i)$, $g_3(t_i)$ and $g_4(t_i)$ equal 1 when the trigram, bigram and unigram probability respectively had a non-zero value for each token $t_i$ in $T$.

Show how you can compute the values for $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ for the equation in Question 3b by using $g_1, g_2, g_3, g_4$. State why the condition stated in your answer to Question 3b is satisfied by your answer.

(4) Consider the following definition for the trigram probability over part of speech tags: $P(t_i \mid t_{i-2}, t_{i-1})$ and the emit probability of a word given a part of speech tag: $P(w_i \mid t_i)$. The part of speech tag definitions are as follows: `bos` (*begin sentence marker*), N (*noun*), V (*verb*), D (*determiner*), P (*preposition*), `eos` (*end of sentence marker*).

| $P(t_i \mid t_{i-2}, t_{i-1})$ | $t_{i-2}$ | $t_{i-1}$ | $t_i$ |
|---|---|---|---|
| 1 | D | N | eos |
| 1 | bos | bos | N |
| 1 | P | D | N |
| $\frac{1}{2}$ | bos | N | N |
| $\frac{1}{2}$ | bos | N | V |
| 1 | V | D | N |
| 1 | V | V | D |
| $\frac{1}{3}$ | N | V | D |
| $\frac{1}{3}$ | N | V | V |
| $\frac{1}{3}$ | N | V | P |
| $\frac{1}{2}$ | N | N | V |
| $\frac{1}{2}$ | N | N | P |
| 1 | N | P | D |
| 1 | V | P | D |

| $P(w_i \mid t_i)$ | $t_i$ | $w_i$ |
|---|---|---|
| 1 | D | an |
| $\frac{2}{5}$ | N | time |
| $\frac{2}{5}$ | N | arrow |
| $\frac{1}{5}$ | N | flies |
| 1 | P | like |
| $\frac{1}{2}$ | V | like |
| $\frac{1}{2}$ | V | flies |
| 1 | eos | eos |
| 1 | bos | bos |

a. The probability of a part of speech tagged sentence

$$s = \texttt{bos/bos}, \texttt{bos/bos}, w_0/t_0, w_1/t_1, \ldots, w_{n-1}/t_{n-1}, \texttt{eos/eos}$$

is given by:

$$P(s) = \prod_{i=0}^{n} P(t_i \mid t_{i-2}, t_{i-1}) \times P(w_i \mid t_i)$$

Using the probability tables given above, compute the probability of the two sentences given below. Which of these two sentences gets the higher probability?

1. `bos/bos`, `bos/bos`, time/N, flies/V, like/P, an/D, arrow/N, `eos/eos`

2. `bos/bos`, `bos/bos`, time/N, flies/N, like/V, an/D, arrow/N, `eos/eos`

b. A hidden markov model is defined as a set of states $\mathbf{s} = (s_0, \ldots, s_m)$ for some fixed, finite value of $m$ where each state $s_i$ can emit an output symbol $w_i$ with probability $P(w_i \mid s_i)$ and each state $s_{i+1}$ can be reached from a state $s_i$ with probability $P(s_{i+1} \mid s_i)$. Provide a hidden markov model (*hmm*) that maps the trigram probabilities $P(t_i \mid t_{i-2}, t_{i-1})$ shown in the table above to transition probabilities $P(s_{i+1} \mid s_i)$ in your hmm. First define a mapping between the trigrams and the set of states in your hmm. Then define the transition probabilities. You don't need to provide the emit probabilities $P(w_i \mid s_i)$.

   You can either draw the hmm graphically as a state machine graph with the transition probabilities on the arcs *or* you can provide a tabular representation of the transition probability $P(s_{i+1} \mid s_i)$. You do not need to show transitions with zero probability or any states that are not useful in representing the trigam probabilities.