(Graded, open-book) Pop Quiz IV (**30 points**)

CS 753 Automatic Speech Recognition
(Scores will not count towards final grade)

October 30, 2017

Name: _____ (Mandatory)

1 ASR evaluations

ASR systems are typically evaluated using the word error rate (WER) metric. For a test set with N utterances $s_1, \ldots, s_N$, WER of predictions $s'_1, \ldots, s'_N$ is defined as $\sum_{i=1}^{N} \frac{\Psi(s_i, s'_i)}{|s_i|}$ where Ψ is the edit distance function with words forming the alphabet and $|s|$ denotes the total number of words in a sequence s.

(A) Define a recurrence for computing Ψ. **[4 points]**

(B) In this problem we consider modifying the edit distance metric used in defining WER to account for compound words such as "fat-free", "pre-lunch", etc. Recall that Ψ allows unit-cost edit operations insert, delete and substitute. Define a new edit distance metric Ψ^* which allows the following two additional unit-cost operations: join which replaces two consecutive words by a compund word derived by joining them (e.g., "fat free" $\rightarrow$ "fat-free") and its inverse operation split which replaces a compound word by its two constituent words. For two sequences s, s' it is possible that $\Psi^*(s, s') < \Psi(s, s')$. For example, if $s =$ "dry-clean the dress" and $s' =$ "dry clean the dress." we have $\Psi(s, s') = 2$ (corresponding to, say, one substitution "dry-clean" $\rightarrow$ "dry" and one insertion of "clean"), but $\Psi^*(s, s') = 1$ (i.e. "dry-clean" $\rightarrow$ "dry clean"). How should the recurrence in the algorithm for computing Ψ be modified to compute Ψ^*? **[8 points]**

(C) If an ASR system misrecognizes a test utterance corresponding to the word sequence "singing in the rain" as either "singing in the reign" or "singing in the universe", both hypotheses would contribute the same error of $\Psi = 1$ to the original WER metric. The fact that one of the misrecognized words "reign" is acoustically much more similar to the original word "rain" than the other misrecognized word "universe" is not taken into account by the WER metric. Can you think of a way in which Ψ can be modified to address this issue? **[2 points]**

2 Neural end-to-end ASR systems

You've been introduced to two neural models used in end-to-end ASR sytems: Connectionist Temporal Classification (CTC) and sequence-to-sequence models with attention (S2S-a). Here are three dimensions along which both these approaches differ. Briefly describe how they differ.

(A) Alignment between input and output symbols is monotonic. **[2 points]**

(B) Hard alignments vs. soft alignments between input/output steps. **[2 points]**

(C) Conditional independence between predictions at different time steps. **[2 points]**

3 The mandatory HMM question

Consider an HMM $\lambda = (A, B)$ with a sequence of hidden states Q, a sequence of observations O, transition probabilities $a_{ij} = \Pr(q_t = j|q_{t-1} = i)$ and emission probabilities $b_j(o_t) = \Pr(o_t|q_t = j)$. The forward probability, $\alpha_t(j)$ and backward probability $\beta_t(j)$ for an observation sequence $\{o_1, \ldots, o_T\}$ of length T are defined as follows:

$$\alpha_t(j) = \Pr(o_1, \ldots, o_t, q_t = j) \tag{1}$$
$$\beta_t(j) = \Pr(o_{t+1}, \ldots, o_T|q_t = j) \tag{2}$$

Compute the following posterior probabilities using the quantities $\alpha_t(j)$, $\beta_t(j)$, a_{ij} and $b_j(o_t)$:

(a) $\Pr(q_{t+1} = k|q_t = j, o_1, \ldots, o_T)$ **[5 points]**

(b) $\Pr(q_{t-1} = i, q_t = j, q_{t+1} = k|o_1, \ldots, o_T)$ **[5 points]**

4 Revision topics

Which of these topics would you like to see revised during the last class? Please write down your order of preference (e.g. A > B > C > D > E > F), starting with the most preferred topic.

(A) WFST-based algorithms for ASR

(B) HMM-based acoustic models in ASR (incl. tied-state models)

(C) Language modeling

(D) Search & decoding

(E) DNN/RNN based ASR models

(F) Discriminative training