CSCI 5832 Spring 2008 Final

Name: _____

On my honor, as a University of Colorado at Boulder student, I have neither given nor received
unauthorized assistance on this work. _____ .

1. **(5 points)** What is your favorite kind of food?

2. **(10 points)** Assume you've been given the task of finding all instances of the verb *eat* in
    a large corpus. That is, you need to find all instances of the verb *eat* in all of its various
    surface forms. You start out with the following regular expression `/eat/`. Characterize
    three kinds of errors (give examples) that will arise with this pattern. Suggest solutions to
    each of the error types you cite.

3. **(5 points) True or False:** The intuition underlying the Good-Turing method is to treat
    unseen events (0 count events) like singleton events (events with a count of 1).

4. Assume in the context of HMMs, that *S* refers to a sequence of states, *O* refers to a
    sequence of observations, and that *M* is a particular HMM model. *Name* the algorithms
    that compute the following values:

    a) **(5 points)** argmax P(S|O,M)

    b) **(5 points)** P(O| M)

5. **(15 points)** Many languages make extensive use of diacritics (i.e., accent marks, as in *naïve*). These marks can provide information critical to the meaning of a word (as in *resume* vs. *résumé*). Unfortunately, many language sources (the Web, newswires, close-captions, etc.) delete or lose this information. Describe the design of a system that restores this accent information to a stream of text without accents. That is, given an unaccented stream of tokens as input, return an appropriately accented stream of tokens as output. Include the basis for the system's design, what kind of resources would be needed, and how you would go about creating the system. **Use the back. Hint: this is not a question about morphology. No morphological processing need be applied in your answer.**

6. **(10 points)** Describe the purpose of the computation occurring in the **If statement** in the following code for PCKY. That is, describe what its doing and why its being done.

```
function PROBABILISTIC-CKY(words,grammar) returns most probable parse
                                        and its probability
   for j←from 1 to LENGTH(words) do
      for all { A | A → words[j] ∈ grammar }
         table[j−1, j,A]←P(A→ words[j])
      for i←from j−2 downto 0 do
         for k←i+1 to j−1 do
            for all { A | A → BC ∈ grammar,
                        and table[i,k,B] > 0 and table[k, j,C] > 0 }
               if (table[i,j,A] < P(A → BC) × table[i,k,B] × table[k,j,C]) then
                  table[i,j,A]←P(A → BC) × table[i,k,B] × table[k,j,C]
                  back[i,j,A]←{k,B,C}
   return BUILD_TREE(back[1, LENGTH(words), S]), table[1, LENGTH(words), S]
```

7. **(10 points)** Give an event-oriented FOL representation for the following sentence.

   John promised Mary he would fly to Miami

8. **(5 Points) True or False:** The *Lesk Algorithm* for WSD requires a machine-readable dictionary as a resource.

9. **(15 Points)** Consider the following line from a recent New York Times article.

> **Charlton Heston**, who appeared in some 100 films in his 60-year acting career but who is remembered chiefly for his monumental, jut-jawed portrayals of Moses, Ben-Hur and Michelangelo, died Saturday night at his home in **Beverly Hills**.

Describe *two kinds* of features that would be useful for detecting and classifying the relation between the named entities in bold in this text. Discuss any potential issues or difficulties that would be relevant to the features you select.

10. **(15 points)** Assume that you are building a *phrase-based* statistical system to translate **from** *Spanish* **to** *German*. Describe three critical sources of information (tables of probabilities, etc.) that your system would need. Be specific; make your answers specific to this particular language pair. By this, I don't mean you need to know Spanish or German. Just describe the nature of the tables using these languages as examples.