CS160 - Homework 5
Due: Wednesday Oct. 9, Before class

For this homework, you will be reading a recent conference paper, answering some questions about the paper and then providing a review of the paper. The paper is relatively short (8 pages), however, it is fairly dense. It will likely take a few passes over the paper to understand it. I don't expect you to understand every detail/equation, but you should understand the paper at a high-level and be able to clearly answer the questions below. Given that this is your homework for the week, I'm expecting it to take approximately the same amount of time as your previous homeworks, so do allocate some time for this.

Because of the density of the paper, I've provided a "reading dictionary" that tries to give a short definition for some of the terms that may be new/confusing. I **strongly** encourage you to read over the paper soon and then come talk to me if there are any questions/confusions.

The paper is:
Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal and Yi Liu. 2007. In *ACL*.
http://www.stefanriezler.com/PAPERS/ACL07.pdf

1. Reading dictionary:

   - **Question Answering**: The field of question answering is a subfield of IR and tackles the specific problem of how to answer user questions, for example "Who was the 5th president?". The most common approaches attempt to extract that information from a corpus (e.g. the web or FAQ articles)
   - **Statistical machine translation (SMT)**: Approaches for doing language translation (e.g. translate Chinese text into English text) based on statistical models.

- **Language model**: A language model estimates how likely a segment of text, like a sentence, is to be an english sentence. For example, a language model would score ”Anteater’s bad is” worse than ”Anteater’s are bad”.
- **Translation model**: A probabilistic model of how language gets translation from one language to another.
- **Syntax, Syntactic information**: Syntax refers to the rules for how sentences or constructed, i.e. grammar of a language. Syntactic information includes things like part of speech (noun, verb, preposition), etc.
- **Word/Phrase alignment**: SMT systems train on bilingual data consisting of sentences in two languages that are translations. The alignment between two such sentences is the phrase/word correspondence between words/phrases in the sentences (which may not be one-to-one). For example, if the word “dog” would be aligned with the word “perro”.
- **n-best translations**: For SMT, we can output not only the best translation, but a list of the n best translations for an input.
- **Bilingual phrase table**: A table with a phrase in one language and the corresponding phrase in another language that are likely to be translations. These phrase pairs also have associated probabilities, i.e. the probability that they are translations.
- **Classifier**: A program that, given an input, determines which of a prespecified set of classes that input belongs to. In the paper, they trainand FAQ classifer, that determines if a web page is an FAQ page or not.

2. (12 points) Questions

   (a) (2 points) The paper refers to the “lexical chasm” and “term mismatch” in the introduction. What does this refer to?

   (b) (2 points) What is the difference between “global” and “local” query expansion?

   (c) (3 points) The basic data point this paper uses is a question-answer pair. Initially, they start with nothing. They do three major steps to get a reasonable data set. What are these steps?

   (d) (3 points) List three things used/discussed in this paper that we have covered in class.

(e) (2 points) How does S_2@20 relate to the precision at 20? Are they the same? If not, how are they different?

3. (15 points) Review

   For your final projects you will be doing a project report. In addition, you will also be asked to review drafts of your classmates reports and give feedback. To give you some practice at doing this, you will be providing a review of this paper. Your review should include the following (make explicit sections for each bullet):

   - Summarize the work in 1 paragraph
   - What problem are the authors trying to solve?
   - What are this work's main contributions?
   - Is this work technically sound? (i.e., How good is the solution? How convinced are you that it works?)
   - Comment on the quality of the writing.
   - Did the paper convince you that they accomplished what they set out to do?

   For your review, be brief and to the point. Back your claims/comments using specific examples from the paper. Imagine that your comments will be read by the authors of the paper, so make your comments as helpful as possible.

4. (Optional) Course feedback

   We're about a third of the way through the course and I'd like to get your feedback on the course so far and where you'd like it to go. If you have a few minutes, please answer the questions below that you have feedback concerning. Feel free to hand this in on another piece of paper without your name on it or you can slide it under my door, etc. I will take all the feedback very seriously and try and alter things accordingly.

   - Are you happy with the course so far?
   - What is one way the course could be improved?
   - What has been your favorite part of class?
   - What has been your least favorite part of class?
   - How could the lectures be improved?

- How is the workload for the course?
- How is the pacing of the class? Too fast? Too slow?
- How is the textbook?
- Are there any topics not on the syllabus you'd like to see covered? Anything you'd not like to see covered?
- Other?