Linguistics 439/539
Fall '14

Final homework

1. Using the Welsh Assembly data, implement a word alignment algorithm that can work from Welsh to English or vice versa, and test it on at least one file in the data set. Turn in the code and answer the following questions:

   (a) How does your algorithm work?

   (b) What words in Welsh does your system predict for:

   - member
   - question
   - government
   - questions

   (c) What words in English does your system predict for:

   - cenedlaethol
   - Cymru
   - iawn
   - Gymru

   (d) How might you improve your system?

2. Imagine an IR query *Middle Feast* (incorrect; *East*, not *Feast*). Implement an SNLP system that will recover from this and return the right results. Test your system on the tagged Brown corpus. Make sure you give results and explain clearly how your system works.

3. Using the Brown corpus, select a content word and a non-content word and compute the following quantities:

   (a) document frequency

   (b) collection frequency

   (c) IDF

   (d) RIDF

   (e) α and β of the K mixture

4. **539 students only**: Using the first 10,000 words of the (untagged) Brown corpus, write code to reduce the bigram space so that words are clustered into only 100 clusters. Your code should return a $2 \times n$ matrix (that can be plotted) of mutual information loss values and number of clusters.

Things to remember:

1. This is due by email by *noon* on **Dec. 16**. (You can email it to me before then if you prefer.)
2. Code must be in the form of a working/runnable m-file that you email to me.
3. Prose questions can be answered as a commented section in an m-file or in a pdf file. Microsoft Word documents are *not* acceptable.
4. The total prose can be no more than *five* double-spaced pages.
5. Remember to comment your code so I can figure out what you're doing conceptually and programmatically.
6. Remember: nothing late. Do not wait until the last minute to do this.
7. You may certainly talk to each other about this and other assignments, but everyone must turn in their own work. (If you do talk to others, I need to see evidence that you are doing more than listening and writing down what others say.)