Assigned: March 22, 2004
Due: April 5, 2004

This assignment covers collocations, word sense disambiguation, and statistical machine translation.

1. (5 points) Collocations
   Exercise 5.4 from Manning/Schuetze, page 176

2. (5 points) Collocations (cont’d)
   Exercise 5.5 from Manning/Schuetze, page 176

3. (35 points) Word Sense Disambiguation using the dictionary method.

   (a) (2 points) Using the Nutch interface at http://lada.si.umich.edu:9090/aquaint/en/search.html, pick five news stories that contain the word “plant” as a noun. Submit printouts of these documents.

   (b) (2 points) Identify the context in which these instances occur. The context should be small enough to be manageable and still large enough to ensure non-trivial overlap with the dictionary definitions.

   (c) (2 points) Identify all senses related to “plant” (as a noun) in wordnet. Let this be set WN.

   (d) (2 points) Identify all senses related to “plant” (as a noun) in the Merriam-Webster dictionary (http://www.m-w.com). Let this be set MW.

   (e) (2 points) Using both the WN and MW sets, identify the “correct” sense of each of the five occurrences of plant.

   (f) (5 points) For the WN sense set, apply the Lesk dictionary-based algorithm to identify the score of each sense. Use either cosine or overlap as the similarity metric. What is the most likely sense for each of the five instances of “plant”? What is the accuracy of this method? Compare with a baseline method.

   (g) (5 points) The same question but using the MW sense set.

   (h) (5 points) Discuss the results from the previous two questions. Explain why you are getting the particular accuracy that your algorithm obtained (why not higher, why not lower)?

   (i) (5 points) Now compute a similarity matrix between all senses in the combined WN+MW sense set. Are there any natural groupings of senses (either within or across sense sets)?
(j) (5 points) Describe a similarity method that you think would give you a better accuracy. You may or may not use the results from the previous item. How well does your new similarity method work compared to cosine or overlap? Discuss why you are or you are not getting an improvement.

4. (10 points) Word Sense Disambiguation papers
Write a short (two pages total) review of these two papers by David Yarowsky: “Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French.” and “Unsupervised Word Sense Disambiguation Rivaling Supervised Methods”. Discuss in particular the “one sense per discourse, one sense per collocation” hypotheses.

5. (5 points) Machine Translation and Language Models
Exercise 13.8 from Manning/Schuetze, page 492.

6. (40 points) Machine Translation
(a) (5 points) Find three instances of Web-based parallel corpora (i.e., sets of documents that appear in parallel in more than one language). Print a sample page from each.
(b) (5 points) Download one sample parallel pair of documents from one of the parallel corpora. Clean it up and convert it to a one-sentence-per-line format (in both languages). Number the sentences in each document (in language A and language B). Note that the number of sentences in the two documents may not be the same.
(c) (5 points) For each word in language A, create a list of the sentences in which it appears. E.g., “cat: 1, 5, 10” means that the word “cat” (or “Cat” or “CAT”) appears in sentences 1, 5, and 10 in the English document.
(d) (5 points) Pick all words in each language that occur in 2, 3, or 4 sentences each (that is, words like “the” should not be on your list). If you get more than ten words with such frequencies in each language, choose any ten among them. The number of words in each language doesn’t need to be the same.
(e) (10 points) Using a modification of the cosine metric, compute the similarities between all pairs of words \((v, w)\) such that \(v\) comes from language A and \(w\) comes from language B. Note that cosine alone will not be very helpful unless the sentences in the two languages are completely aligned. In general, they are not. You have to invent a measure of similarity that is high for likely translations even if the sentences are not perfectly aligned.
(f) (10 points) Using the result of the previous item, indicate the most likely translation of each word in language A (chosen among all candidates from language B). Discuss how well (or poorly) this algorithm works.

You can do this last question either by hand or by using a program. If you choose to do it manually, limit yourself to documents that have no more than approximately 15 sentences each (that is, don’t spend days doing simple computations). On the other hand, if you write a program, it should work on arbitrarily long documents.