

# CS 421 – Natural Language Processing – Spring 2010

## Homework 2 - Part B

### General Information

**Deadline (Part B): Friday March 12, midnight**

**Worth:** 16 points (out of 1000 total for the class)

**What and how to hand it in:** all the required files in a single zip file

### 1 The LoPar parser (12 points)

Download the LoPar parser from the course web site; you find it under Assignments / Homework 2b. This is also the parser we will use for the class project.

LoPar runs under Linux (or Solaris, if anybody still has access to Suns). If you don't have Linux on your own machine, you can install it in your directory on the CS Department server – oscar is the Linux server (your TA has tried it, and the parser runs without a snag)

The parser comes with a precompiled binary for Linux or Solaris. Just unpack the appropriate archive file and the parser is ready to use.

In the data folder, you will find the following three files:

- example.gram - the grammar file.
- example.lex - the lexicon file.
- example.txt - input file with test sentence.

Note: All grammar rules in *example.gram* are of the form:  $LHS \rightarrow RHS_1 RHS_2 \dots RHS_n$ , where  $LHS$  is the left hand side non-terminal (parent category) and  $RHS_i$  are the right hand side non-terminals (daughter categories). All non-terminals should start with an UPPERCASE letter. As for the lexicon (*example.lex*), each line contains an entry for a word (in LOWERCASE) followed by a tab character and then a list of all possible POS categories for that particular word. Ignore capitalizations and punctuation marks for this exercise. The parser input file *example.txt* must have one-word-per-line format, i.e. each line corresponds to

one word in the sentence. The end of a sentence is marked by an empty line. More than one sentence in the input file should be separated by the empty line.

To run the parser in symbolic mode (i.e., no probabilities), you can use these parameters:

```
./lopar -symbolic -trees <grammar file> -l <lexicon file> <input file>
```

With the `-trees` option the parser returns all parse trees.

Your task for this exercise is to augment the grammar and lexicon files (*example.gram* and *example.lex*) so that the following sentences are parsed by LoPar (include these sentences in *example.txt* and run the parser):

- Davis won a gold medal at the Vancouver Olympics
- Which races did Lindsey Vonn win?
- The Italian skater who won gold in Torino lost in Vancouver

## 2 The Stanford parser (4 points)

The Stanford parser is a statistical parser available at <http://nlp.stanford.edu/software/lex-parser.shtml>. There's an online demo at this same URL, click on "ONLINE".

Try the three sentences above (about the Olympics) with the Stanford parser. You will get three results back: the input sentence tagged for POS, a parse, and dependencies. Take the parse and disregard the dependencies.

## 3 What to submit

Please upload 4 files via the class website: the modified files for LoPar, *example.gram*, *example.lex*, and *example.txt*; and a result file that contains the output parses from LoPar and from the Stanford parser, for the three sentences about the Olympics. The result file will also include a comparison between the two sets of parses: discuss any differences between them. Additionally, answer the following question: for LoPar, how many different sentences can your tiny grammar and lexicon for English generate? you don't need a precise number, only an estimate.

Please follow the link under Assignments / Homework 2B to upload your file. Remember to package all your files in a single .zip or .tar file, and upload that, not individual files.