CS224/Ling 237 Homework #4

Due Wed, 29 May 2001, 5pm

(20 points total)

1. (2 points) Information Extraction. Suppose that one wanted to learn acronyms as extracted facts from documents. So, the learned relation would have two fields, acronym and expansion, as in:

U.S.A.	United States of America
IR	Information Retrieval

Present two (significantly different) extraction rules that would be useful for learning acronyms (if you're having trouble, look at a few research papers, and see how and where acronyms are introduced). These extraction rules could either be something that would be handcoded (perhaps as a regular expression), or they could be instances of a general extraction pattern that you would hope to learn using a machine learning method. For the purposes of this problem, it is sufficient just to state them as examples. You will probably find it helpful to look at some of the examples of extraction rules in the Muslea paper 'Extraction Patterns for Information Extraction Tasks: A Survey', and you might want to borrow the notations that one or more of those systems use in presenting your answer – though do tell us which system's notation you are using.

2. (6 points) Consider a simplified version of an ambiguous sentence example that we used at the beginning of the quarter:

   Fed raises interest rates greatly

   A possible grammar that will be able to parse this sentence is:

S	$\rightarrow$	NP VP	NP	$\rightarrow$	*Fed*
VP	$\rightarrow$	V NP	NP	$\rightarrow$	*raises*
VP	$\rightarrow$	VP Adv	NP	$\rightarrow$	*interest*
VP	$\rightarrow$	V	NP	$\rightarrow$	*rates*
NP	$\rightarrow$	NP NP	Adv	$\rightarrow$	*greatly*
V	$\rightarrow$	*raises*	V	$\rightarrow$	*interest*
V	$\rightarrow$	*rates*			

   We would now like to parse this sentence

   (a) The grammar as shown isn't quite in Chomsky normal form. Turn it into a CNF grammar.

(b) Make a well-formed substring table (parse triangle) for this sentence, and then parse it using the CKY algorithm, filling in all the cells in the triangle with the correct constituents (it isn't necessary to include further annotation as to how they were built up).

(c) The above grammar does not capture that some of the nouns are plural and some singular, and that some of the verbs are forms (like *interest*) which would be used with plural subjects, and others like *rates* would be used with a singular subject. Furthermore, the grammar also doesn't capture that all three of the verbs shown are normally only used transitively (with a required NP object after the verb). Change the grammar so it captures all this information (but is still a CNF grammar).

(d) Parse the sentence again using the CKY algorithm and the new grammar, and again show the completed parse triangle (without further annotation being necessary).

3. (4 points) Recall the phrase structure grammar we used in class:

S	→	NP VP	N	→	*cats*
VP	→	V NP	N	→	*claws*
VP	→	V NP PP	N	→	*people*
NP	→	NP PP	N	→	*scratch*
NP	→	N	V	→	*scratch*
NP	→	e	P	→	*with*
PP	→	P NP			

This grammar includes all the nasties: a left-recursive rule, a unary rule, a more-than-binary rule, and an empty category. We will explore using it in a chart parser to parse sentences. We would suggest drawing the chart in the style of a parse triangle, except this time numbering starting from 0 instead of 1 and treating the numbers as fenceposts. That is, for the sentence *I like brie*, the cell (0,2) will hold chart edges for *I like*, and the cell (2,3) will hold chart edges for *brie.* The parse triangle cell entries will include both passive edges (complete constituents) and active edges (incomplete dotted rules).

(a) Consider parsing this grammar with a bottom up chart parser on the sentence *Cats with claws scratch.* Show the chart that results.

(b) Repeat this exercise parsing top down. Show the chart that results. What differences are there in the work that was done?

4. (3 points) In chapter 11 of M&S, we suggested that PCFGs have a bad bias towards using nonterminals with few expansions. Suppose that one has as a training corpus the treebank given below, where '$n\times$' indicates how many times a certain tree appears in the training corpus. What PCFG would one get from the treebank (using MLE as discussed in the text)? What is the most likely parse of the string 'a a' using that grammar. Is this a reasonable result? Was the problem of bias stated correctly in chapter 11? Discuss.

$$\left\{ \begin{array}{lcclcclcclcclcc}
 & \multicolumn{2}{c}{\mathrm{S}} & & \multicolumn{2}{c}{\mathrm{S}} & & \multicolumn{2}{c}{\mathrm{S}} & & \multicolumn{2}{c}{\mathrm{S}} & & \multicolumn{2}{c}{\mathrm{S}} \\
 & \multicolumn{2}{c}{\wedge} & & \multicolumn{2}{c}{\wedge} & & \multicolumn{2}{c}{\wedge} & & \multicolumn{2}{c}{\wedge} & & \multicolumn{2}{c}{\wedge} \\
10\times & \mathrm{B} & \mathrm{B}, & 95\times & \mathrm{A} & \mathrm{A}, & 325\times & \mathrm{A} & \mathrm{A}, & 8\times & \mathrm{A} & \mathrm{A}, & 428\times & \mathrm{A} & \mathrm{A} \\
 & | & | & & | & | & & | & | & & | & | & & | & | \\
 & a & a & & a & a & & f & g & & f & a & & g & f
\end{array} \right\}$$

5. (2 points) Syntax and Logical Semantics I. As well as *the overpriced car runs*, we can also say *the overpriced car does run.* Starting from the grammar fragments shown in the handout, We can introduce *does* as a word of category Aux and add an additional rule VP $\rightarrow$ Aux VP. Determine a semantic representation for *does*, semantic annotations for the new syntax rule, and show a parse tree complete with semantics showing this rule used to give a correct semantic representation for *Kathy does respect Fong.*

6. (3 points) Syntax and Logical Semantics II. If *woman* and *telescope* are given lexical entries like *car*, how many semantic readings are there for *Kathy saw the woman with a telescope*? Give the parse trees, complete with semantic forms. How do the generated semantic forms not match the semantic distinctions that we have generally assumed for the multiple parses of sentences such as this?