CS224N/Ling 237 Homework #6
PCFGs and Probabilistic Parsing

Due: Wed, May 12 2004

1. CKY Parsing of PCFGs

   One of the most famous ambiguous sentences of English from early computational linguistics work is:

   Time flies like an arrow.

   Using the ambiguity-causing-things that we have standardly used (part of speech, PP attachment, noun compounding), you should be able to generate *four* parse trees (and, hence, different meanings) for this sentence. If you're having trouble thinking of four, or just if you're interested in the history of this sentence, you can look here:

   http://mitpress.mit.edu/e-books/Hal/chap7/seven2.html

   Assume the following grammar G:

S	→	NP VP	0.8	N	→	*time*	0.5
S	→	VP	0.2	N	→	*flies*	0.3
VP	→	V NP	0.5	N	→	*arrow*	0.2
VP	→	V PP	0.3	V	→	*time*	0.3
VP	→	VP PP	0.2	V	→	*flies*	0.3
NP	→	Det N	0.3	V	→	*like*	0.4
NP	→	N	0.3	P	→	*like*	1.0
NP	→	N N	0.2	Det	→	*an*	1.0
NP	→	NP PP	0.2				
PP	→	P NP	1.0				

   (a) Draw the four parses for the sentence using the original grammar above.

   (b) Generate a version of this grammar in Chomsky Normal Form by performing unary rule removal. Make the probabilities of the rules such that the language generated by the grammar is unchanged (sentences get the same probability). (I.e., work out what to do with rule probabilities when unary removal is done.)

   (c) Draw a CKY parse triangle for this grammar parsing the sentence above, including working out the inside probabilities. What is $P(\textit{Time flies like an arrow}|G)$?

   (d) Draw a CKY parse triangle for this grammar parsing the sentence above, working out Viterbi (maximum) probabilities. What is $\arg\max_t P(t|\textit{Time flies like an arrow}, G)$? Indicate which of the four parses in (a) is the one chosen as most likely.

(e) What would be the effect on the probabilities calculated in the previous part of adding some other preposition, say P $\rightarrow$ *on*, with some reasonable non-zero probability? Discuss in a sentence or two whether this seems linguistically sensible in terms of using PCFGs for disambiguating sentences?

2. PCFG probabilities.

   We are given a PCFG over a grammar in Chomsky Normal Form with start symbol N^1. The grammar has a non-terminal set $\{N^1, \ldots, N^n\}$, terminal set $\{w^1, \ldots, w^V\}$, a set of binary rules of the form $N^i \rightarrow N^j N^k$, and a set of unary rules $N^i \rightarrow w^j$.

   We are also given a sentence w_{1m}, which is parsable by the grammar. Provide an expression for the probability that words w_i and w_j from this sentence, where $i < j$, have a single common ancestor (the start symbol N^1). You can use the definitions of inside and outside probabilities (M&S pp. 392–398).

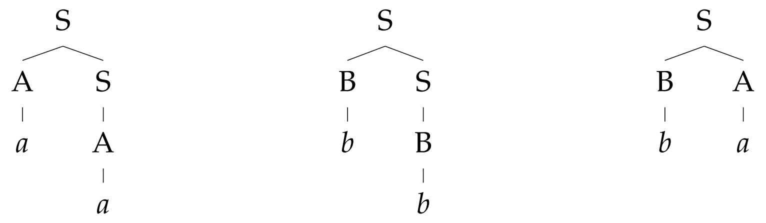
3. Joint and Conditionally Trained PCFG.

   In this problem we will look at parameter estimation for PCFGs by maximizing the joint likelihood and estimation using conditional likelihood.

   Consider the following context free grammar:

   S — start symbol
   $\{S, A, B\}$ — nonterminal set
   $\{a, b\}$ — terminal set
   $\{S \rightarrow A\ S, S \rightarrow B\ S, S \rightarrow B\ A, S \rightarrow A, S \rightarrow B, A \rightarrow a, B \rightarrow b\}$ — rule set

   We are given the following training corpus:



   The first two trees occurred 5 times each and the third tree occurred once.

   (a) Specify the parameters of the PCFG model that maximizes the joint likelihood of the training data.

   (b) What is the error rate of this model on the training set? (Here, we look at the PCFG as a classifier that, for an input sentence , returns its most likely parse tree. If the returned tree is different than the tree that appears for the sentence in the training set, then this is a training set error. The error rate is the fraction of errors.)

   (c) Write out the conditional log-likelihood of the training set using the parameters you found in the previous part. (The conditional probability for a tree t for a sequence of words $w_{1,\ldots,n}$ according to a PCFG grammar is $P(t|w_{1,\ldots,n}) = P(t, w_{1,\ldots,n})/P(w_{1,\ldots,n}) = P(t)/P(w_{1,\ldots,n})$, since the sentence $w_{1,\ldots,n}$ is part of the tree.)

(d) Are there parameter settings for the PCFG that would assign higher conditional log-likelihood to the data? (If so, specify such parameters.)

(e) Are there parameter settings for the PCFG that would achieve lower error rate ? (if so, specify such parameters).