1999

P798
EJB

Solution Notes – NLP 99 Exam Question 1

10 marks for each section

a)
Chart data structure for efficiently encoding a parse forest of syntactic analyses; dag with edges and vertices (= word boundaries); edges represent subanalyses; active chart extension to represent hypotheses; active edge data structure; rule invocation / active edge addition determines top-down / bottom up; agenda determines depth- / breadth- first; no. of edges polynomial in length of input, no of analyses exponential (6mks)
Encoding all analyses does not resolve ambiguity; rule invocation for large grammars may be a big constant overhead; unification and non-determinism problem with extension from CFG to UBGs (4mks)

b)
Local rules for part-of-speech (pos) disambiguation in context; representing rules as n-grams / FSMs (eg. state=pos, transition=legal sequence); dictionary of words and pos assignments; method for computing legal paths (eg. intersection of input as FSM and FSM encoding rules) (5mks)
Usually more than one path thru' FSM will remain, so probabilistic extension to (H)MM (lexical + transition, e.g. bigram, probs.); Viterbi algorithm; lexical probs. esp. difficult to estimate – mle, smoothing, sparse data (5mks)

c)
CFG as encoding of syntactic rules; ambiguity problem (e.g. N –¿ N N, PP attachment); variation + change problem – what is grammatical?; probabilistic CFG = language model, formal defs (p, A –¿ alpha, for all A=NT sum to 1, p(d) = prod p(r in d) etc); ranking derivations to resolve ambiguity; inducing grammar from data (CNF CFG); estimating probs. – mle, smoothing (6mks)
Independence assumptions mean many distinct derivations will have indistinguishable probabilities (e.g. N –¿ N N); use of non-CNF PCFG means shorter derivs will be too strongly preferred; lack of lexical info or easy way to integrate biggest flaw as words resolve most structural ambiguity; grammar induction using PCFG leads to full coverage but arbitrary derivations (4mks)

d)
Planning: initial state-¿goal – decomposition into elementary tasks; preconditions, dependencies, sequencing; plans and speaker goals; speech acts; plan selection as a form of defeasible inference / abduction; forward chaining for interpreting text relative to plans; egs. resolving reference, bridging inferences, etc.; dynamic flexibility vs. scripts (6mks)
Problems of capturing and representing large amounts of knowledge required; plan selection in a realistic database etc and consequent domain-dependence and lack of scalability, portability, etc. (4mks)