COMP90042 Web Search & Text Analysis

Workshop Week 10

Zenan Zhai

May 14, 2019

University of Melbourne

Outline

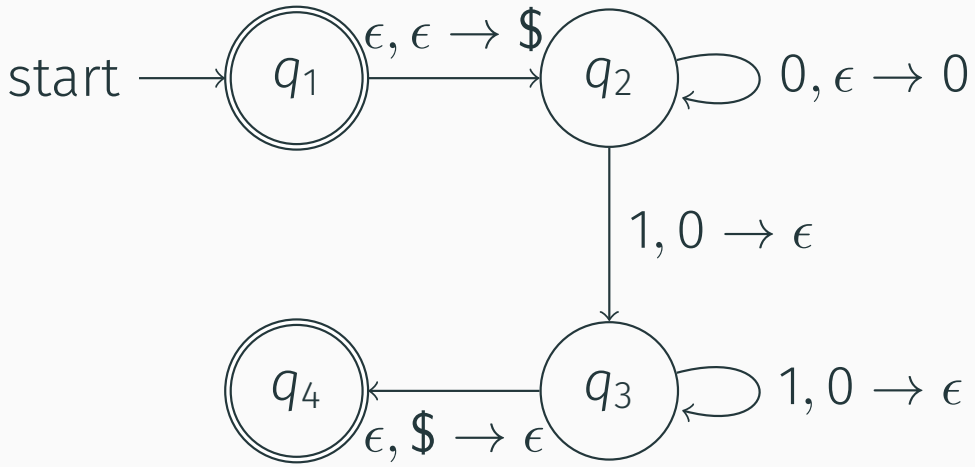
Computation Theory

- Pushdown automaton
- Context-free grammar

Chart parsing

- CYK
- Probabilistic CYK

Pushdown automaton



A pushdown automaton is a 6-tuple $(Q, \Sigma, \Gamma, \delta, q_0, F)$ such that:

1. Q Set of all states
2. Σ Input alphabet
3. Γ stack alphabet
4. q_0 start states
5. F Set of final states
6. $\delta : Q \times \Sigma_\epsilon \times \Gamma_\epsilon \to \mathrm{P}(Q \times \Gamma_\epsilon)$ Transitions

Context-free grammar

A context-free grammar is a 4-tuple (V, Σ, R, S) where

1. V is a finite set called variables,
2. Σ is a finite set, disjoint from V, called terminals,
3. R is a finite set of rules, with each rules being a variable and a string of variables and terminals, and
4. $S \in V$ in the start variable

A context-free grammar is in Chomsky normal form if every rule is of the form

1. $A \to BC$
2. $A \to a$

where a is any terminal and A, B and C are any variables (except start variable).
In addition, we allow $S \to \epsilon$, where S is the start variable

CYK

```
for j ← 1 to Length(words) do
    for all {A|A → word[j] ∈ Grammar} do
     |  table[j − 1, j] ← table[j − 1, j] ∪ A
    end
    for i ← from j − 2 downto 0 do
        for k ← i + 1 to j − 1 do
            for all {A | A →  BC and B ∈ table[i, k] and C ∈ table[k, j]}
             do
             |  table[i, j] ← table[i, j] ∪ A
            end
        end
    end
end
```

- Why do we need to convert CFGs to Chomsky Normal Form?
- What are i, j, k representing?
- Time complexity of CYK?

CYK - Exercise

Given following CFG

- $S \rightarrow NP\ VP$
- $VP \rightarrow V\ NP \mid NP\ PP$
- $PP \rightarrow P\ NP$
- $V \rightarrow\ ''saw'' \mid ''walked''$
- $NP \rightarrow\ ''John'' \mid ''Bob'' \mid Det\ N \mid Det\ N\ PP$
- $Det \rightarrow\ ''a'' \mid ''an'' \mid ''the'' \mid ''my''$
- $N \rightarrow\ ''man'' \mid ''cat'' \mid ''telescope'' \mid ''park''$
- $P \rightarrow\ ''on'' \mid ''by'' \mid ''with''$

Parse the following sentence using CYK

- ”a man saw john”
- ”an park by Bob walked an park with Bob”
- ”park by the cat with my telescope”

Probabilistic CYK

A sentence may have several valid parsing trees

- Assign probs. to CFGs , then use product of them as probs. of the parsing tree.

```
for j ← 1 to Length(words) do
    for all {A | A → word[j] ∈ Grammar} do
        table[j − 1, j, A] ← P(A → words[j])
    end
    for i ← from j − 2 downto 0 do
        for k ← i + 1 to j − 1 do
            for all
              {A | A → BC ∈ grammar, and table[i, k, B] > 0 and table[k, j, C] > 0} do
                if (table[i, j, A] < P(A → BC) × table[i, k, B] × table[k, j, C]) then
                    table[i, j, A] ← P(A → BC) × table[i, k, B] × table[k, j, C]
                    back[i, j, A] ← {k, B, C}
                end
            end
        end
    end
end
```

CYK V.S. Viterbi

Question 6: Markov Models and Grammars [6 marks]

a) Hidden Markov models (HMMs) describe a probability distribution over both words w_i and tags t_i. State the formulation of the joint probability, $p(w_1, w_2, \ldots, w_N, t_1, t_2, \ldots, t_N)$, for a first order hidden Markov model, and use your formulation to describe two of the modelling assumptions. [2 marks]

b) A first order hidden Markov model can be expressed with a probabilistic context-free grammar (PCFG), as follows:

S	$\rightarrow t\ X_t$	[_____]	for all tags t
X_t	$\rightarrow t'\ X_{t'}$	[_____]	for all tag pairs t, t'
X_t	$\rightarrow$ </s>	[_____]	for all tags t
t	$\rightarrow w$	[_____]	for all tags t and words w

where S is the start symbol, X_t are a special non-terminal symbols (for each tag t), </s> is a special terminal denoting the the end of the sentence, and each rule has a probability denoted in square brackets, []. Fill in the empty boxes with the relevant parts of the probability formulation of the first order hidden Markov model (from your answer to part 1). [2 marks]

c) Using the above grammar illustrate the equivalence between the CYK algorithm for PCFG parsing and the Viterbi algorithm for the HMM. You might want to show an example sentence to aid your explanation. [2 marks]

from exam paper 2016