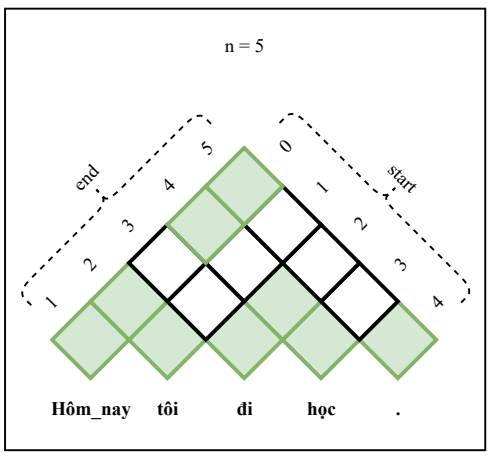
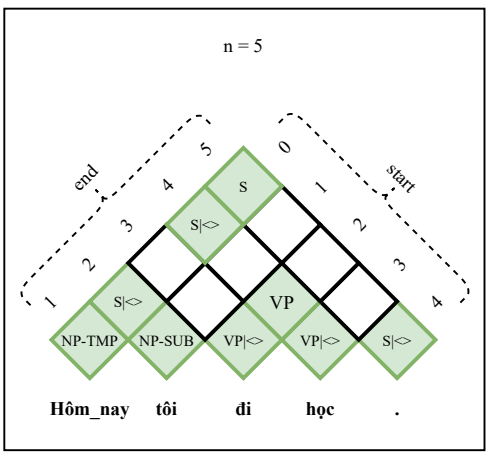
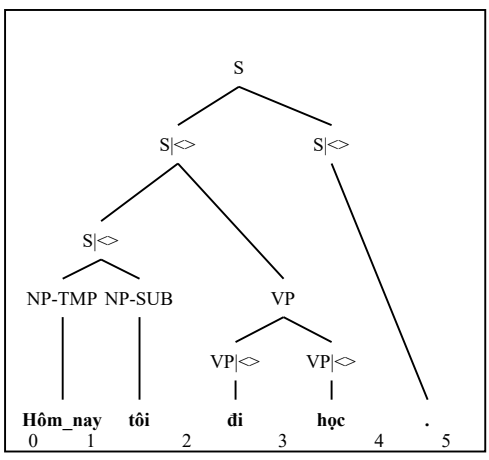


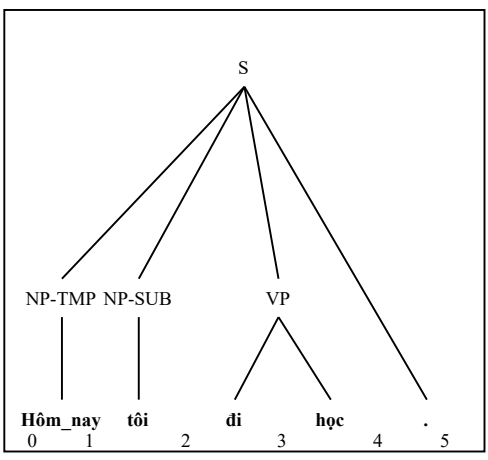
Bước 1: Các ngữ đoạn được tìm bởi thuật toán CKY.



Bước 2: Gán nhãn cú pháp CNF cho các ngữ đoạn.



Bước 3: Cây CNF.



Bước 4: Cây CFG.

(S(NP-TMP(Hôm_nay))(NP-SUB(tôi))(VP(đi)(học))(.))

Bước 5: Cây CFG ở định dạng dấu ngoặc.

Hình 1. Năm bước phổ biến để dự đoán một cây cú pháp thành tố CFG ở định dạng dấu ngoặc.

Yêu cầu bài tập: Cài đặt thuật toán CKY khi biết các điểm số của các ngữ đoạn được dự đoán từ mô hình Span-Based Neural Constituency Parsing

Trong bài tập này, bạn sẽ cài đặt lại thuật toán CKY với định dạng đầu vào như sau:
- Dòng đầu tiên chứa một số nguyên dương **n** đại diện cho số từ trong câu đầu vào (**n** có giá trị tối đa là **100**).
- Dòng thứ hai chứa một chuỗi gồm **n** từ cách nhau bởi một ký tự khoảng trắng. Biết rằng, không tồn tại các ký tự mở ngoặc "(" và đóng ngoặc ")" trong một từ.
- Cuối cùng, **(n(n+1)/2)** dòng tiếp theo, mỗi dòng chứa số nguyên thứ nhất là chỉ số **start**, số nguyên thứ hai là chỉ số **end** của một ngữ đoạn, thứ ba là một số thực không âm có giá trị không âm đại diện cho điểm số của ngữ đoạn **[start, end]**, chuỗi không chứa khoảng trắng nằm ở cuối chính là nhãn cú pháp CNF của ngữ đoạn **[start, end]**.

Và định dạng đầu ra như sau:
- Cây CFG ở định dạng dấu ngoặc được hiển thị trên một dòng.

Lưu ý quan trọng trong bài tập này, đó là, bạn phải sử dụng thuật toán CKY để **tìm cây cú pháp có tổng điểm các ngữ đoạn lớn nhất** mà **không cần quan tâm đến nhãn cú pháp CNF**. Chúng ta mặc định rằng các nhãn cú pháp CNF cung cấp trong đầu vào được coi dự đoán đúng. Sau khi tìm được cây cú pháp có tổng điểm lớn nhất, bạn sẽ tiến hành gán nhãn CNF cho các ngữ đoạn đó. Cuối cùng, bạn phải chuyển cây CNF về CFG ở định dạng dấu ngoặc để in ra kết quả cuối cùng. Ngoài ra, trong bài tập này, các nhãn từ loại (part-of-speech tags) đã được loại bỏ. Đây là hướng tiếp cận của phần lớn công cụ phân tích cú pháp trong những năm gần đây. Bạn có thể xem Hình 1 ở phía trên và đầu vào mẫu phía dưới để nắm rõ hơn các bước để làm bài tập này.

Khi chuyển đổi cây CNF về CFG, bạn cần lưu ý hai trường hợp sau:
- Ngữ đoạn **[start, end]** có nhãn kết thúc bằng "|<>" chỉ thị các cây con của ngữ đoạn **[start, end]** được coi là các cây con của "cha của ngữ đoạn **[start, end]**" khi chuyển từ CNF sang CFG, sau đó chúng ta loại bỏ ngữ đoạn **[start, end]** đi. Đây là cách để đưa cây nhị phân (CNF) về cây nhiều nhánh (CFG) bất kể ban đầu chúng ta chuyển từ cây CNF sang CFG bằng cách đệ quy trái hay đệ quy phải.
- Nhãn CNF của ngữ đoạn có chứa "**::**" chỉ thị cây một nhánh chứa các nút theo thứ tự được tách bởi "**::**". Ví dụ, nhãn CNF "**X::Y::Z**" được coi như luật CFG "**X → Y → Z**", nhãn CNF "**X::Y**" được coi như luật CFG "**X → Y**".

Ngoài ra, bạn có thể thấy trong bài tập này, có thể xuất hiện cả tiếng Việt và tiếng Anh. Nhưng bạn đừng bận tâm, bởi vì nhãn cú pháp CNF được cung cấp trong đầu vào được coi là đã dự đoán đúng. Bạn có thể tìm thêm mô tả các nhãn cú pháp tiếng Việt tại đường dẫn https://vlsp.org.vn/vlsp2021/eval/parsing hoặc tiếng Anh tại đường dẫn http://surdeanu.cs.arizona.edu//mihai/teaching/ista555-fall13/readings/PennTreebankConstituents.html.

Đầu vào mẫu:	Đầu ra mẫu:
5 Hôm_nay tôi đi học . 0 1 1.0000001192092896 NP-TMP 0 2 0.9998076558113098 S|<> 0 3 3.294167254352942e-05 S|<> 0 4 1.0 S|<> 0 5 1.0 S 1 2 1.0000001192092896 NP-SUB 1 3 4.483771931518277e-08 S|<> 1 4 0.00019242330745328218 SBAR::S 1 5 2.1107127157193872e-09 S 2 3 1.0000001192092896 VP|<> 2 4 0.9999671578407288 VP 2 5 2.8640911864385998e-08 VP|<> 3 4 1.0000001192092896 VP|<> 3 5 5.17452747317293e-12 VP|<> 4 5 1.0 S|<>	(S(NP-TMP(Hôm_nay))(NP-SUB(tôi))(VP(đi)(học))(.))
5 He enjoys playing soccer . 0 1 0.9999998211860657 NP 0 2 7.203940185718238e-05 S|<> 0 3 2.740061518125003e-06 S|<> 0 4 0.9999880194664001 S|<> 0 5 1.0 S 1 2 0.9999998211860657 VP|<> 1 3 0.0009769656462594867 VP|<> 1 4 0.9999252557754517 VP 1 5 1.1614187314989977e-05 VP 2 3 0.9999998211860657 VP|<> 2 4 0.9990225434303284 S::VP 2 5 3.6240527379050036e-07 S::VP 3 4 0.9999998211860657 NP 3 5 3.138477921993399e-08 S|<> 4 5 1.0 S|<>	(S(NP(He))(VP(enjoys)(S(VP(playing)(NP(soccer)))))(.))