

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH



BÁO CÁO THỰC NGHIỆM

Vietnamese Constituency Parsing Tree

Môn: CS221.O12.KHCL - XỬ LÝ NGÔN NGỮ TỰ NHIÊN

GVHD: cô Nguyễn Thị Quý

Các thành viên:

1. Huỳnh Võ Ngọc Thanh - 21520449
2. Hồ Yến Nhi - 21520380
3. Nguyễn Trần Hoài Bảo - 21520618
4. Lê Văn Trường - 21522733

TP.HCM, ngày 28 tháng 12 năm 2023

1 Giới thiệu chung:

Phân tích cú pháp theo thành phần (constituency parsing) là quá trình phân tích dựa trên một ngữ pháp cấu trúc từ (phrase structure grammar).

Thông tin về cú pháp đóng một vai trò quan trọng trong nhiều ứng dụng như dịch máy, trích xuất thông tin, trả lời câu hỏi, v.v. Từ năm 2015 trở đi, các mô hình phân tích dựa trên học sâu đã mang lại những thành công mới cho vấn đề này, nhưng chủ yếu là đối với các ngôn ngữ phổ biến như tiếng Anh và tiếng Trung.

Trong bài này, nhóm sẽ trình bày về *Vietnamese Constituency Parsing Tree*, nhằm giúp cho mọi người có thể hiểu rõ hơn về cấu trúc cũng như ngữ pháp của tiếng Việt:

- Input: là một câu, ví dụ như "Nam làm bài tập."
- Output: là một cây cú pháp có dạng:
(S (NP (N Nam)) (VP (V làm) (NP (N bài_tập))))

2 Bộ dữ liệu:

- Sử dụng bộ dữ liệu VLSP là một file txt gồm khoảng 10,00 câu ở dạng bracketed-tree.

- Một sample dữ liệu lấy từ VLSP:

((S-TTL(NP-SUB(N-H Đất)(A nghèo)) (VP(V-H trở_mình)) (. .)))

- Để download bộ dữ liệu và xem cách thiết kế tập nhãn cú pháp, hướng dẫn gán nhãn, các bạn có thể truy cập theo link sau: **link**

- Một vài chú ý khi gán nhãn cụm từ:

Phần tử trung tâm của một cụm từ (ngữ đoạn) có các thuộc tính sau:

- + Nó là yếu tố mang tất cả các thuộc tính ngữ pháp của ngữ đoạn.
- + Nó là yếu tố duy nhất của ngữ đoạn có thể có quan hệ ngữ pháp và ngữ nghĩa vượt ra ngoài ngữ đoạn
- + Các yếu tố khác của ngữ đoạn chỉ có quan hệ phụ thuộc trực tiếp hay gián tiếp với trung tâm ngữ đoạn mà thôi (chứ không có bất cứ quan hệ gì ra ngoài phạm vi ngữ đoạn).

3 Thực nghiệm:

Ý tưởng chính: sử dụng thư viện *Stanza* có tích hợp sẵn trình xử lý dữ liệu và gán nhãn từ loại với trình phân tích cú pháp theo thành phần trên dữ liệu VLSP bằng mô hình pre-trained PhoBert-large.

Stanza là một thư viện xử lý ngôn ngữ tự nhiên mã nguồn mở, được phát triển bởi đội ngũ nghiên cứu từ Stanford University.

Đầu tiên, tiến hành cài đặt stanza với pip:

```
1 !pip install stanza
```

Cài đặt thành công sẽ xuất hiện như sau:

```
Installing collected packages: emoji, stanza  
Successfully installed emoji-2.9.0 stanza-1.7.0
```

Import vào các thư viện cần thiết:







```
1 from nltk import Tree  
2 import torch  
3 import stanza
```

Dùng stanza để tạo pipeline và tải về các nguồn tài nguyên cần thiết cho tiếng Việt, bao gồm các trình tokenize, pos, và constituency.


```
1 vi_tree = stanza.Pipeline(lang='vi', model_dir=save_model,  
    processors='tokenize, pos, constituency', treebank=True, use_gpu=False,  
    verbose=None)
```

```
=====
| Processor      | Package                               |
|-----|-----|
| tokenize       | vtb                                   |
| pos            | vtb_phobert-large                   |
| constituency   | vlsp22_phobert-large               |
=====
```


Ta được các folder lưu các trọng số cho từng pre-trained model:

	backward_charlm
	constituency
	forward_charlm
	pos
	pretrain
	tokenize


Tokenization biến đổi dữ liệu văn bản thành định dạng mà mô hình có thể hiểu và xử lý. File trọng số của model **tokenize**:

 **vtb.pt**

POS là quá trình gán nhãn từ loại cho mỗi từ trong một câu. Cụ thể, mỗi từ được xác định là một loại từ ngữ nhất định như danh từ, động từ, tính từ, trạng từ, giới từ, và nhiều loại từ khác. File trọng số của model **pos**:

 **vtb_phobert-large.pt**

Constituency thường được sử dụng để mô tả các thành phần ngôn ngữ trong cấu trúc câu. File trọng số của model **constituency**:

 **vlsp22_phobert-large.pt**

4 Kết quả thực nghiệm:

Nhập vào câu input tiếng Việt:

```
vi_sentence = 'Tôi là sinh viên khoa Khoa học máy tính.'
```

Kết quả demo:

(ROOT (S (NP (Pro Tôi)) (VP (V là) (NP (N sinh viên) (NP (N khoa) (N Khoa) (V học) (NP (N máy) (N tính)))))) (. .)))

Ngoài ra, mình có demo thêm English Constituency Parsing Tree, nếu các bạn có nhu cầu phân tích cú pháp tiếng Anh, mình xin mời các bạn vào đường link sau: **link colab**

5 Nguồn tham khảo:

Adding a new Constituency model

An Empirical Study for Vietnamese Constituency Parsing with Pre-training

Berkeley Neural Parser

Multilingual Constituency Parsing with Self-Attention and Pre-Training