TÌM HIỂU VÀ PHÂN TÍCH CÚ PHÁP CẤU TRÚC NGỮ ĐOẠN

GVHD: Nguyễn Trọng Chỉnh

1. Trịnh Quang Trường 18520393
2. Phạm Xuân Thiên 18520158
3. Tăng Năng Chung 18520536

NỘI DUNG

I. GIỚI THIỆU ĐỀ TÀI
II. TÓM TẮT QUY TRÌNH
III. THU THẬP DỮ LIỆU
IV. TẠO BỘ NGỮ LIỆU
V. CÚ PHÁP CẤU TRÚC NGỮ ĐOẠN
VI. KẾT LUẬN

I. GIỚI THIỆU ĐỀ TÀI

Ứng dụng:

- Hiểu về ý nghĩa, cấu trúc và các mối quan hệ trong câu. Phục vụ cho bài toán **Information Extraction**[1]

Bên cạnh thứ hạng cùng sân nhà King Power, lợi thế của Leicester là quãng thời gian 1 tuần nghỉ ngơi, trong khi MU phải đá tứ kết League Cup rạng sáng 24/12 (thắng Everton 2-0). Còn nhớ mùa giải năm ngoái, MU từng đánh bại Leicester 2-0 ngay tại King Power đúng vào vòng cuối cùng để giành suất dự Champions League, vì vậy đoàn quân do Brendan Rodgers dẫn dắt đang khát khao phục hận.

I. GIỚI THIỆU ĐỀ TÀI

Ứng dụng:

- Mô phỏng toàn bộ cấu trúc câu

  Hệ thống **Kiểm tra Ngữ pháp**.

Grammarly Bad Content Check

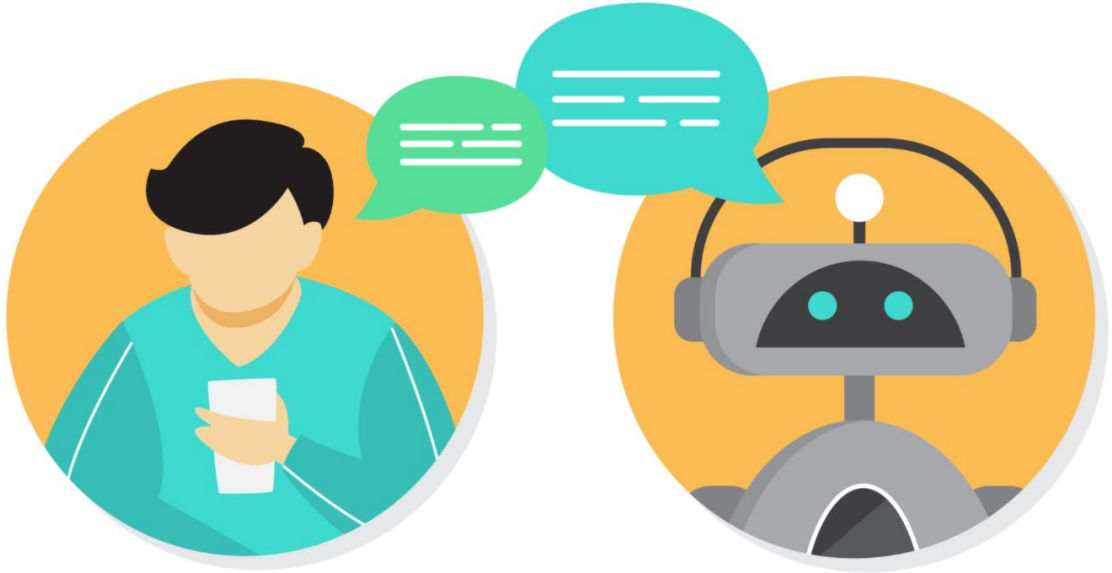
Depends on the package you choose, GoDaddy will provide you with a grate set of tools. You will get several ready made solutions for blogging, e-commerce, client management and forums. All you need is to put the tools you need in to a whole system. Basic features like disk space and bandwidth is good in any package, just choose the type depending in the potential capacity of your project.

With a 99.9% Uptime guarantee, GoDaddy offer a service that just fall below perfect here, and give a confidence boost too anyone who is looking four a reliable hosting provider. With a site scaner, strong virus protection and server monitoring, GoDaddy will help you run a smooth and efficient website with minimal disruption.

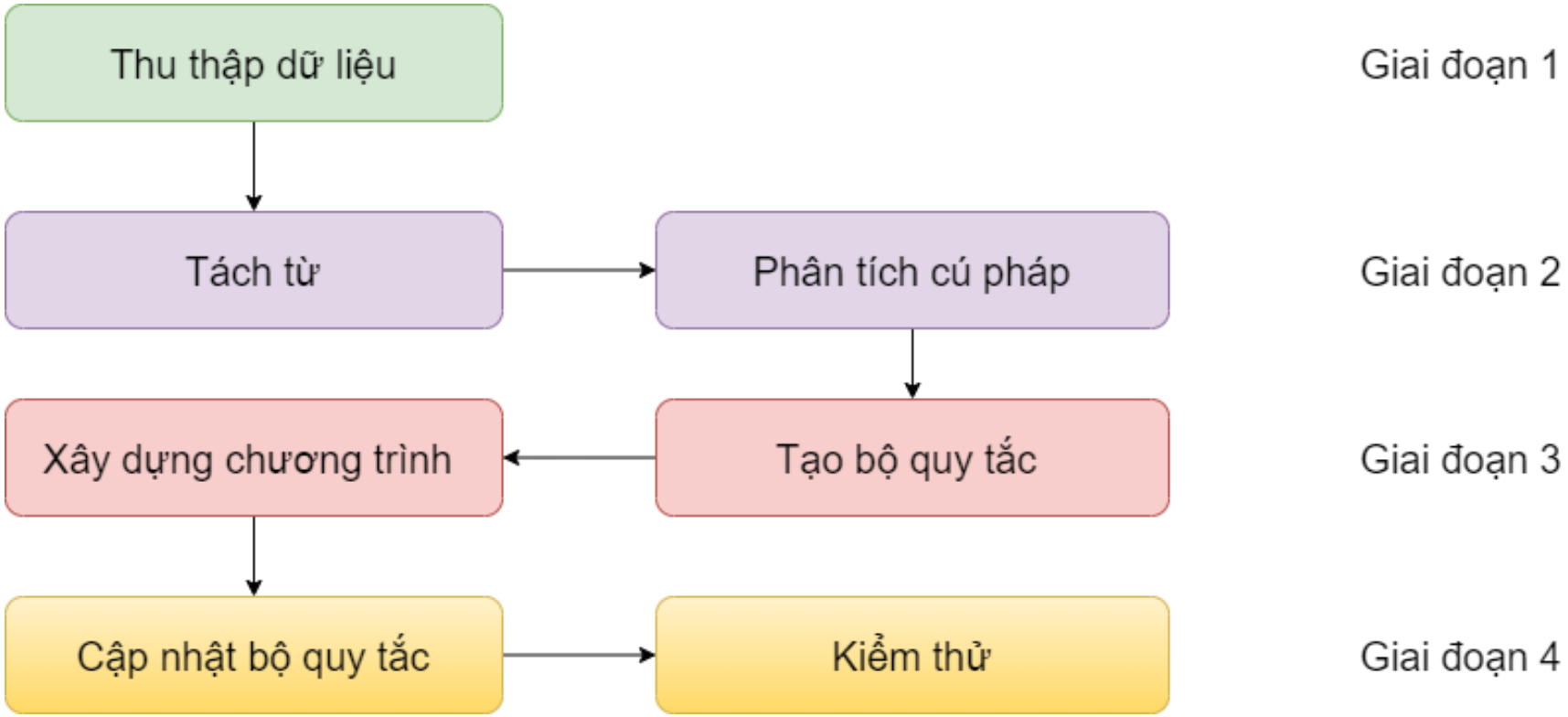
I. GIỚI THIỆU ĐỀ TÀI

Ứng dụng:

- Vai trò biểu diễn trung gian cho các bài toán **Semantic Annalysis, Question Answering**..



II. TÓM TẮT QUY TRÌNH



Hình 4. Sơ đồ quy trình xây dựng bộ dữ liệu

III. THU THẬP DỮ LIỆU

YÊU CẦU:

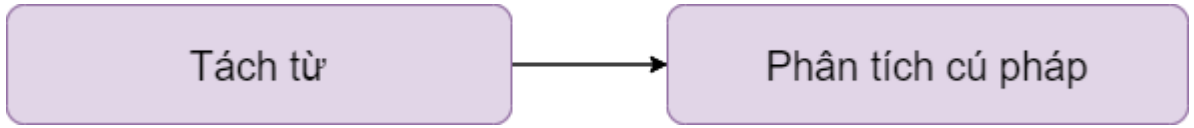
- Thu thập trên ngữ liệu thực tế phục vụ cho bài toán phân tích ngữ pháp cấu trúc ngữ đoạn

THỰC HIỆN:

- Thu thập bộ dữ liệu là tiêu đề bài báo trên trang https://tuoitre.vn/
- Bộ dữ liệu bao gồm : 78 câu
- Tìm hiểu bộ nhãn của guildline VLSP 7.3

IV. TẠO BỘ NGỮ LIỆU

GIAI ĐOẠN 2:

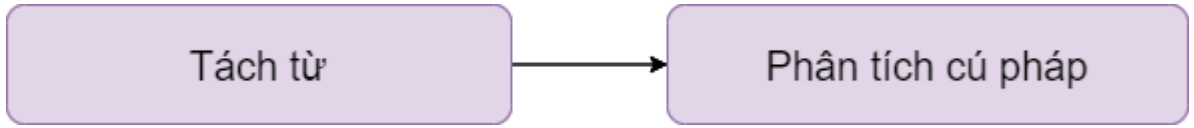


❖ TÁCH TỪ:

Lựa chọn thuật toán: Maximum Matching

Ý tưởng: Sử dụng 1 bộ từ điển chứa tất cả các từ đang xét và sẽ duyệt từ trái qua phải(hoặc ngược lại) và chọn từ dài nhất có thể chọn được nếu không thì sẽ giảm độ dài của từ đó và tiếp tục kiểm tra cho đến khi hết câu.

IV. TẠO BỘ NGỮ LIỆU

GIAI ĐOẠN 2: 

❖ TÁCH TỪ:

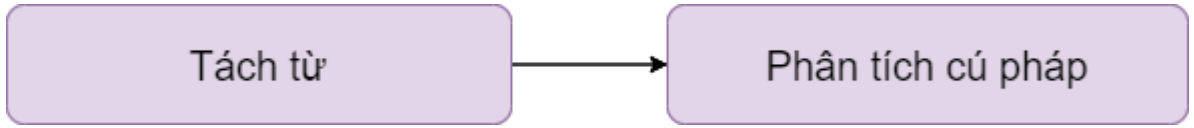
Lựa chọn thuật toán: Maximum Matching

Kết quả:

	Precision	Recall
Maximum Matching	99.583	98.974
VnCoreNLP	95.458	95.315

IV. TẠO BỘ NGỮ LIỆU

GIAI ĐOẠN 2:



❖ PHÂN TÍCH CÚ PHÁP:

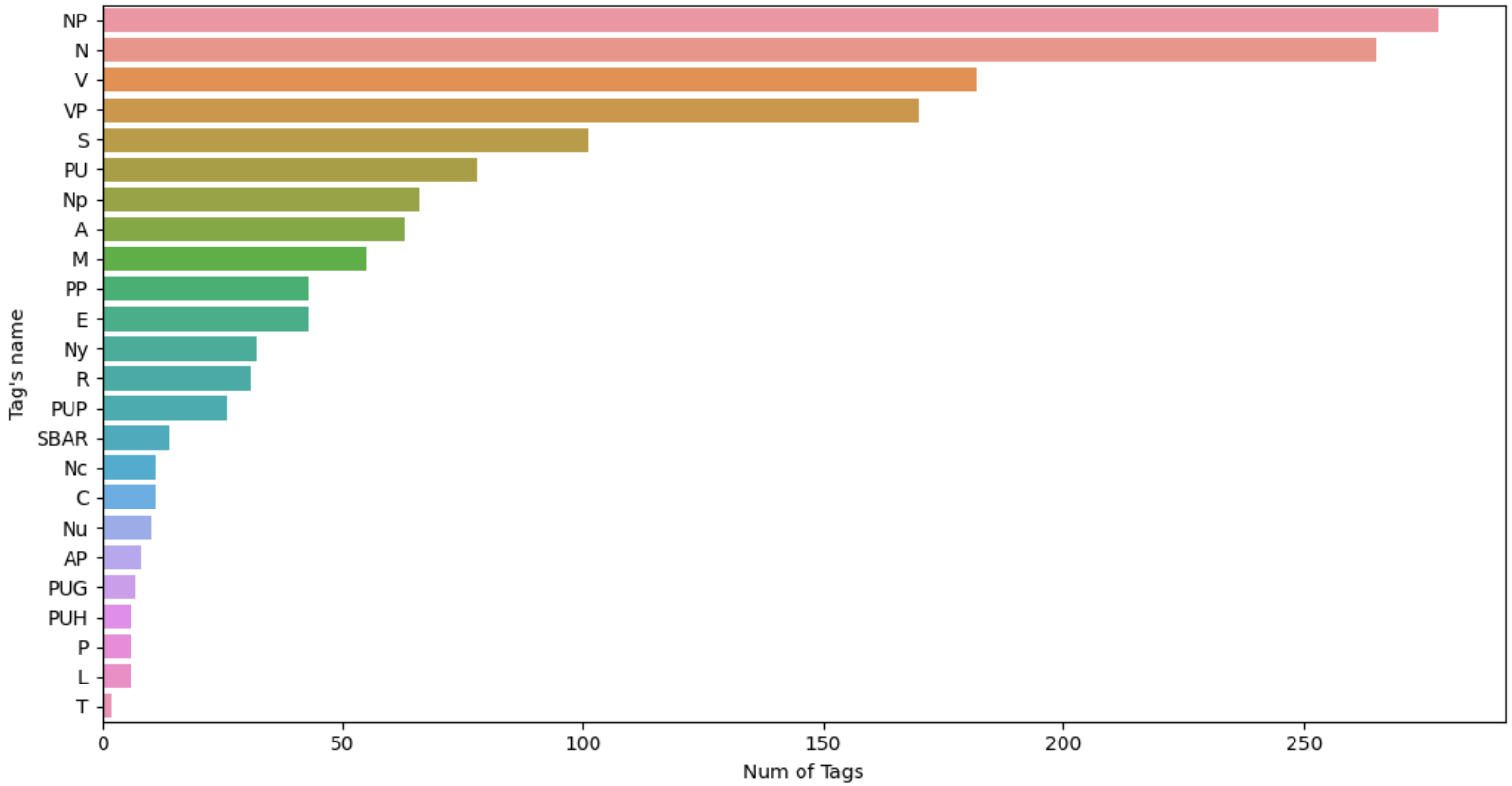
- Tìm hiểu bộ nhãn VLSP 7.3
  - Số nhẫn sử dụng 24 nhãn : NP, VP, AP, S, PP,…
- Gắn nhãn thủ công
  - Thảo luận xử lý các nhập nhằng:

    Ví dụ: (VP (V thu_hồi) (NP (A hơn) (NP (M 6.000) (NP (Nu ha) (N đất)))))

IV. TẠO BỘ NGỮ LIỆU

GIAI ĐOẠN 2:

Tách từ → Phân tích cú pháp



V. CÚ PHÁP CẤU TRÚC NGỮ ĐOẠN

GIAI ĐOẠN 3: Tạo bộ quy tắc → Xây dựng chương trình

❖THUẬT TOÁN CKY()

Cho văn phạm G = (N, Σ, P, S) có dạng chuẩn Chomsky và một câu s, thuật toán CKY phân tích cú pháp câu s như sau:

- Giả sử s = x1 x2 .. xn với xi là từ thứ i trong câu s

- Gọi T(i, j, X) là cấu trúc của cụm từ có nhãn X bắt đầu bởi từ thứ i và kết thúc bởi từ thứ j

- Nếu tồn tại T(i, k, Y1) và T(k, j, Y2) và có luật sản sinh X → Y1 Y2 thì tồn tại T(i, j, X).

- Thuật toán kết thúc nếu xác định được T(1, n, S)

V. CÚ PHÁP CẤU TRÚC NGỮ ĐOẠN

GIAI ĐOẠN 3: Tạo bộ quy tắc → Xây dựng chương trình

❖ĐẶC ĐIỂM CỦA THUẬT TOÁN CKY

- Phân tích cú pháp theo văn phạm phi ngữ cảnh (CFG) .
- Tìm kiếm theo chiến lược Bottom-up.
- Là thuật toán quy hoạch động.
- Các luật sản sinh phải được chuyển về dạng chuẩn Chomsky – CNF (Chomsky Normal Form).

V. CÚ PHÁP CẤU TRÚC NGỮ ĐOẠN

GIAI ĐOẠN 3: Tạo bộ quy tắc → Xây dựng chương trình

❖TẠO BỘ QUY TẮC:

- Xây dựng cấu trúc cú pháp(Grammar) và cấu trúc từ vựng(Lexicon) theo chuẩn CNF(Chomsky Normal Form)
- Một văn phạm CFG G = (N, Σ, P, S) được gọi là có dạng chuẩn Chomsky nếu mỗi luật sản sinh trong P có một trong hai dạng sau: $X \rightarrow Y1\ Y2$ với $X,Y1,Y2 \in N \cup \Sigma$.

  $X \rightarrow Y$ với $X \in N$, $Y \in \Sigma$.

V. CÚ PHÁP CẤU TRÚC NGỮ ĐOẠN

GIAI ĐOẠN 3: Tạo bộ quy tắc → Xây dựng chương trình

❖ XÂY DỰNG BỘ GRAMMAR VÀ BỘ LEXICON

```
1  S -> VP PU
2  S -> S PU
3  S -> NP VP
4  S -> Np VP
5  PP -> E N
6  PP -> E NP
7  AP -> A A
8  AP -> R A
9  AP -> A AP
10 AP -> AP A
11 AP -> AP AP
12 AP -> A C
```

```
21 NP -> M N
22 NP -> M NP
23 NP -> M Nu
24 NP -> M Np
25 NP -> A NP
26 NP -> A N
27 NP -> N A
28 NP -> N AP
29 NP -> N N
30 NP -> N Np
31 NP -> N V
32 NP -> N M
```

```
70 VP -> A V
71 VP -> V AP
72 VP -> V NP
73 VP -> V R
74 VP -> V N
75 VP -> V A
76 VP -> V PP
77 VP -> R V
78 VP -> PU VP
79 VP -> R VP
80 VP -> V VP
81 VP -> VP V
```

Một số luật ở trong bộ Grammar

```
20 A -> khó
21 A -> quý
22 AP -> trái_phép
23 N -> băng_giá
24 R -> hàng
25 P -> Tôi
26 Nc -> người
27 V -> gắn
28 V -> kẹp
29 V -> phạt
30 N -> xe_máy
31 Np -> Nguyễn_Ngọc_Tuấn
```

```
35 M -> 2020
36 N -> nghị_quyết
37 N -> bản_thân
38 N -> Bạn
39 N -> đường
40 V -> chối
41 N -> liều
42 V -> khởi_công
43 Np -> Quốc_hội
44 A -> mới
45 V -> ghi_nhận
46 V -> trình_chiếu
```

Một số từ có trong Lexicon

V. CÚ PHÁP CẤU TRÚC NGỮ ĐOẠN

GIAI ĐOẠN 3: Tạo bộ quy tắc → Xây dựng chương trình

- INPUT : Tập Grammar, Lexicon, Sentence
- OUTPUT : Parsed sentence

❖ CÁC BƯỚC XÂY DỰNG CHƯƠNG TRÌNH

- B1: Tách từ
- B2: Xét Lexicon
- B3: Tìm Grammar
- B4: Truy vết cây cú pháp

V. CÚ PHÁP CẤU TRÚC NGỮ ĐOẠN

GIAI ĐOẠN 4: 

❖ CẬP NHẬT BỘ QUY TẮC:

- Sau khi chạy thử chương trình thì phát hiện 1 số trường hợp đặc biệt cần phải xử lý.

  Ví dụ: “Chủ tịch Hà Nội : ‘ Các ứng viên phó chủ tịch đều rất có năng lực . ’ ”
- Từ đó cập nhật thêm các luật.

V. CÚ PHÁP CẤU TRÚC NGỮ ĐOẠN

GIAI ĐOẠN 4: Cập nhật bộ quy tắc → Kiểm thử

❖ KẾT QUẢ KIỂM THỬ

		Precision	Recall
CKY	First_Result	0.6616	0.6509
	Average	0.5858	0.5746
StarfordCoreNLP	First_Result	0.8239	0.8022

V. KẾT LUẬN

- Thuật toán Maximum Matching.

- Thuật toán CKY.

TÀI LIỆU THAM KHẢO

[1] Ming Jiang, Jana Diesner *A Constituency Parsing Tree based Method for Relation Extraction from Abstracts of Scholarly Publications*

[2] Đinh Điền, *Xử lý ngôn ngữ tự nhiên*. NXB ĐHQGHCM, 2006

Thanks for your attention