**KHOA CÔNG NGHỆ THÔNG TIN**

**ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH**

Thực hiện:

20C11035 – Trương Thế Kiệt

20C11040 – Đặng Nhật Minh

BÁO CÁO ĐỒ ÁN CÀI ĐẶT ỨNG DỤNG

NĂM HỌC 2020-2021

**Xử lý ngôn ngữ tự nhiên nâng cao**

**ĐẠI HỌC QUỐC GIA TPHCM**

**BẢNG THÔNG TIN CHI TIẾT NHÓM**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên đề tài:** | **UD07 - Xây dựng ứng dụng web hỗ trợ phân tích cú pháp thành phần cho văn bản tiếng Anh** | | |
| **Số lượng thành viên:** | **2** | | |
| **MSSV** | **Họ tên** | **Email** | **Điện thoại** |
| 20C11035 | Trương Thế Kiệt | truongthekiet709@gmail.com |  |
| 20C11040 | Đặng Nhật Minh | minhdangnhat685@gmail.com |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **BẢNG PHÂN CÔNG & ĐÁNH GIÁ HOÀN THÀNH CÔNG VIỆC** | | | |
| **Người thực hiện** | **Công việc thực hiện** | **Mức độ hoàn thành** | **Đánh giá của nhóm** |
| 20C11035  Trương Thế Kiệt | Xây dựng phần back-end và front-end cho trang web | 80% | 8/10 |
| Visualize cây cú pháp |
| Chuẩn bị và tích hợp data về từ điển và CFG production rules vào giải thuật phân tích cú pháp |
| Viết báo cáo |
| 20C11040  Đặng Nhật Minh | Cài đặt giải thuật phân tích cú pháp: Earley | 80% | 8/10 |
| Xây dựng cây cú pháp từ kết quả giải thuật phân tích |
| Hỗ trợ chuẩn bị data |
| Viết báo cáo |

MỤC LỤC

[I. TỔNG QUAN 2](#_Toc76583718)

[1. Mô tả bài toán 2](#_Toc76583719)

[2. Các thư viện và dữ liệu sử dụng trong dự án 2](#_Toc76583720)

[II. THUẬT TOÁN SỬ DỤNG 3](#_Toc76583721)

[1. Giới thiệu các thuật toán hiện có 3](#_Toc76583722)

[2. Giới thiệu về giải thuật Earley 3](#_Toc76583723)

[3. Giới thiệu giải thuật xây dựng cây cú pháp 5](#_Toc76583724)

[III. CHI TIẾT CÀI ĐẶT 6](#_Toc76583725)

[1. Giới thiệu code và chức năng từng class 6](#_Toc76583726)

[a. Phần cài đặt chính của thuật toán Earley 6](#_Toc76583727)

[b. Phần cài đặt thuật toán xây dựng cây cú pháp 7](#_Toc76583728)

[2. Sơ đồ hoạt động 8](#_Toc76583729)

[IV. CHẠY CHƯƠNG TRÌNH 9](#_Toc76583730)

[1. Các bước cần thực hiện 9](#_Toc76583731)

[2. Kết quả 9](#_Toc76583732)

[V. KẾT LUẬN 11](#_Toc76583733)

[VI. TÀI LIỆU THAM KHẢO 12](#_Toc76583734)

# TỔNG QUAN

## Mô tả bài toán

Phân tích cú pháp là quá trình phân tích một chuỗi những biểu tượng tuân theo các quy tắc của ngữ pháp hình thức, áp dụng trong ngôn ngữ tự nhiên, ngôn ngữ máy tính, và các cấu trúc dữ liệu khác. Trong ngôn ngữ tự nhiên, sau khi phân tích ta thu được một cây cú pháp đại diện cho cấu trúc cú pháp (syntactic structure) của câu theo các cấu trúc ngữ pháp. Việc phân tích cú pháp sẽ mang lại nhiều lợi ích, như trong việc xây dựng một trình biên dịch, nó giúp kiểm tra một giá trị đầu vào có đúng với cú pháp của ngôn ngữ đang xét hay không

Dựa vào kiến thức đã học nhóm sẽ xây dựng một ứng dụng giúp phân tích cú pháp câu với ngôn ngữ tiếng Anh dựa vào các bộ luật và bộ tự vựng của ngôn ngữ.

## Các thư viện và dữ liệu sử dụng trong dự án

|  |  |
| --- | --- |
| **Tên thư viện** | **Mục đích sử dụng** |
| Spring Boot | Xây dựng ứng dụng web |
| CLC Syntactic Tree Viewer | Giúp trực quan hóa cây phân tích cú pháp dưới dạng cây đồ thị |
| Thymeleaf | Cài đặt front-end |

|  |  |
| --- | --- |
| **Tên dữ liệu** | **Mục đích sử dụng** |
| Penn English | Bộ từ vựng |

# THUẬT TOÁN SỬ DỤNG

## Giới thiệu các thuật toán hiện có

* Với hướng tiếp cận theo luật (rule-based) chúng ta có nhiểu thuật toán giúp phân tích cú pháp của một câu không dựa vào ngữ cảnh. Kĩ thuật phân tích cấu trúc được chia làm loại chính:
  + Top Down Parsing
  + Bottom Up Parsing
* Top Down Parsing: cây cú pháp được xây dựng bắt đầu từ gốc (root) sau đó đi dần tới các nút lá (leaf node). Một số thuật toán sử dụng kĩ thuật này: LL Parser, **Earley Parser**, Pratt Parser, Recursive Descent Parser, Parser Combinator
* Bottom Up Parsing: Cây cú pháp được xây dựng dựa từ các node lá tiến dần về các node gốc. Một số thuật toán sử dụng kĩ thuật này: CYK Parser, LR Parser, Generalized LR Parser

## Giới thiệu về giải thuật Earley

* Giải thuật Earley là một trong những giải thuật được sử dụng phổ biến trong việc xây dựng các hệ thống phân tích cú pháp. Giải thuật được công bố vào năm 1970 bởi Jay Earley, sử dụng chiến lược phân tích kiểu trên xuống (top-down search, ngược với giải thuật CKY vốn là kiểu bottom-up) để phân tích cú pháp của câu bằng cách sử dụng ngôn ngữ phi ngữ cảnh (context free grammar).
* Sử dụng quy hoạch động để cài đặt, độ phức tạp trung bình O(n^3)
* Được ứng dụng rất phổ biến trong thực tế: trình biên dịch (compiler), môi trường tích hợp để phát triển ứng dụng (Integrated Development Environment - IDE), … do có thể phân tích được tất cả các loại ngôn ngữ phi ngữ cảnh

**Ý tưởng thuật toán**

* Thuậ toán Earley sử dụng các luật sản sinh (production rules) và siêu ký hiệu (meta-symbol) để biểu diễn quá trình hoạt động
* Mỗi trạng thái (state) gồm một luật sản sinh và một siêu ký hiệu được đặt ở một vị trí ở vế phải (right hand side) trong luật đó để thể hiện quá trình phân tích luật đó, cấu trúc chứa luật và siêu ký hiệu gọi là dotted rule. Ví dụ: NP -> Det @ Nominal, ta có luật sản sinh NP -> Det Nominal và siêu ký hiệu @ đặt trước Nominal trong vế phải, nghĩa là luật này đã phân tích xong phần Det, và đang phân tích Nominal
* Với một câu N từ, thuật toán Earley sẽ duyệt qua từng từ để điền các trạng thái vào N+1 bảng (chart) với trạng thái được tạo ra bởi các hành động: predict, scan, và complete
* Predict: tạo ra trạng thái mới để thể hiện những mong đợi trong quá trình phân tích, được gọi khi gặp một trạng thái với siêu ký hiệu đứng trước một non-terminal
* Scan: khi gặp một trạng thái với siêu ký hiệu đứng trước một terminal, scan được gọi để kiểm tra xem từ đang xét trong câu có đúng với mong đợi theo luật hay không
* Complete: được gọi khi gặp trạng thái (s1) mà siêu ký hiệu đứng ở cuối cùng (nghĩa là ta đã hoàn thành luật này) để quay lui lại những trạng thái trước đây đang chờ trạng thái s1 này hoàn tất để tiếp tục được phân tích, complete sẽ tạo ra trạng thái mới bằng cách sao chép trạng thái cũ và dời siêu ký hiệu sang bên phải

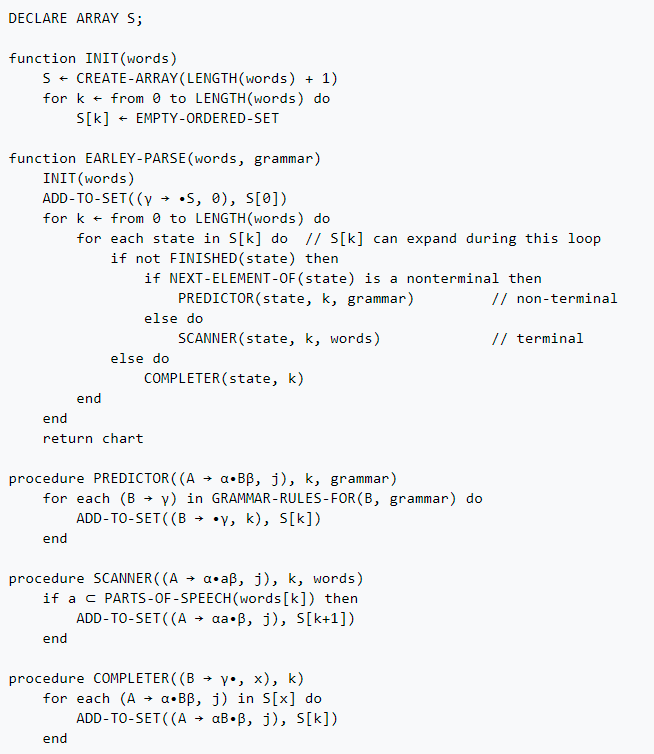


Figure 1 Mã giả thuật toán Earley

## Giới thiệu giải thuật xây dựng cây cú pháp

Sau khi chạy thuật toán Earley cho câu N từ, ta thu được N+1 bảng với các trạng thái

Text

Description automatically generated

Figure 2 Kết quả thuật toán Earley

Với hành động complete, trong mỗi trạng thái ta lưu lại các trạng thái sinh ra nó.

Tìm trong bảng cuối cùng các trạng thái có vế trái (left hand side) là ký hiệu bắt đầu (S). Từ đó ta quay lui về các trạng thái trước để vẽ ra cây cú pháp

# CHI TIẾT CÀI ĐẶT

## Giới thiệu code và chức năng từng class

### Phần cài đặt chính của thuật toán Earley

Text

Description automatically generated

Figure 3 Phần cài đặt chính của thuật toán Earley

### Phần cài đặt thuật toán xây dựng cây cú pháp

Text

Description automatically generated

Figure 4 Phần cài đặt thuật toán xây dựng cây cú pháp

## Sơ đồ hoạt động

Diagram

Description automatically generated

# CHẠY CHƯƠNG TRÌNH

## Các bước cần thực hiện

1. Để tiến hành chạy được chương trình cần chỉnh sửa một số cấu hình trong file **application.yml**

server:

port: 7080 #port run BE API

host: http://localhost # host hiện tại đang chạy back end API

grammar:

is-read-default: false #Đọc giá trị grammar đã được đóng gói trong jar hay không

grammar-file-path: ./Grammar.txt #khi giá trị field is-read-default là true thí không cần set field này.

1. Một ví dụ cho giá trị của file **Grammar.txt**:

S>NP VP

VP>V NP

NP>Conj NP

NP>N NP

NP>N

NP>Pron

NP>N Conj VP

1. Thư mục hiện tại sẽ bao gồm các file:

application.yml

grammar.parser-0.0.1-SNAPSHOT.jar

Grammar.txt

1. Thực hiện mở terminal và chạy dòng lệnh sau:

**java -jar grammar.parser-0.0.1-SNAPSHOT.jar**

1. Sau khi đã start thành công mở browser và truy cập vào đường dẫn: localhost:7080 để sử dụng app.

## Kết quả

Kết quả sau khi start chương trình thành công với config như trên, và nhập giá trị text input là:

**she has breakfast and drink milk**

Kết quả nhận được là:

Chart

Description automatically generated

# KẾT LUẬN

* Nhóm đã đạt được mục tiêu ban đầu là xây dựng một ứng dụng hỗ trợ phân tích cú pháp cho ngôn ngữ tiếng Anh bằng cách áp dụng những kiến thức đã học về xử lý ngôn ngữ tự nhiên, đặc biệt là Context free grammar và các giải thuật phân tích cú pháp theo luật (rule-based)
* Đã cài đặt được một chương trình tương đối hoàn chỉnh có thể sử dụng được. Tuy nhiên tính hiệu quả của ứng dụng trong việc phân tích cú pháp tiếng Anh vẫn còn hạn chế do phụ thuộc vào bộ luật, nhiều câu có cấu trúc phức tạp không phân tích được
* Có thể áp dụng trong việc phân tích các ngôn ngữ phi ngữ cảnh khác

# TÀI LIỆU THAM KHẢO

[1] Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing: Second Edition, Daniel Jurafsky & James H. Martin

[2] Earley parser algorithm <https://en.wikipedia.org/wiki/Earley_parser>

[3] Parsing techniques <https://www.guru99.com/syntax-analysis-parsing-types.html>

[4] Parsing guide <https://tomassetti.me/guide-parsing-algorithms-terminology/>