BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC KINH TÊ - KỸ THUẬT CÔNG NGHIỆP

**KHÓA LUẬN TỐT NGHIỆP**

TÊN ĐỀ TÀI KHÓA LUẬN:

**TÌM HIỂU XỬ LÝ NGÔN NGỮ TỰ NHIÊN VÀ VIẾT CHƯƠNG TRÌNH KIỂM TRA CẤU TRÚC CÂU TRONG TIẾNG ANH**

**Ngành đào tạo: Công nghệ thông tin**

**Mã số ngành: 7480201**

**Họ và tên sinh viên: Hà Gia Kiệt, Trịnh Đức Kiên**

Người hướng dẫn khóa luận tốt nghiệp

**Th.S Lê Thị Kiều Oanh**

**Hà Nội - 2025**

LỜI CAM ĐOAN

Tôi xin khẳng định rằng toàn bộ nội dung trình bày trong luận văn này là kết quả của quá trình nghiên cứu và làm việc nghiêm túc của cá nhân tôi, được thực hiện dựa trên sự hướng dẫn của các thầy cô, cùng với sự nỗ lực tìm tòi và học hỏi. Luận văn đã được xây dựng trên cơ sở khoa học, thực tiễn và sử dụng các nguồn tài liệu tham khảo phù hợp.

Tôi cam đoan rằng nội dung luận văn không sao chép từ bất kỳ công trình nghiên cứu nào khác, ngoại trừ những tài liệu, dữ liệu, hoặc ý tưởng được tôi tham khảo và trích dẫn đầy đủ theo quy định trong danh mục tài liệu tham khảo. Nếu phát hiện bất kỳ vi phạm nào liên quan đến vấn đề bản quyền hay tính trung thực, tôi xin hoàn toàn chịu trách nhiệm trước quy định của nhà trường cũng như pháp luật.

Tôi chân thành cảm ơn sự giúp đỡ của những người đã hỗ trợ tôi trong quá trình thực hiện nghiên cứu này.

LỜI CẢM ƠN

Kính thưa Quý Thầy/Cô Khoa Công nghệ Thông tin, trong suốt hành trình học tập tại Khoa, đặc biệt là giai đoạn thực hiện luận văn tốt nghiệp, chúng em đã may mắn nhận được sự quan tâm, hướng dẫn tận tình và sự hỗ trợ quý báu từ Quý Thầy/Cô. Với tất cả lòng kính trọng và biết ơn, chúng em xin được bày tỏ lời tri ân sâu sắc đến:

Cô Lê Thị Kiều Oanh: Người đã dành trọn tâm huyết, thời gian và sự tận tâm để trực tiếp hướng dẫn chúng em trong quá trình nghiên cứu. Sự chỉ dạy tận tình, những lời khuyên sâu sắc và sự động viên kịp thời của Cô đã tiếp thêm động lực lớn lao, giúp chúng em vượt qua mọi khó khăn để hoàn thiện luận văn. Cô không chỉ là người dẫn dắt mà còn là nguồn cảm hứng, giúp chúng em khám phá và phát huy tiềm năng của bản thân.

Quý Thầy/Cô Khoa Công nghệ Thông tin: Với tri thức sâu rộng và kinh nghiệm phong phú, Quý Thầy/Cô đã trang bị cho chúng em nền tảng chuyên môn vững chắc, khơi nguồn đam mê và tiếp thêm nghị lực để chúng em tự tin trên con đường chinh phục tri thức. Chúng em vô cùng biết ơn những điều kiện thuận lợi và sự hỗ trợ tận tình mà Khoa đã dành cho chúng em trong suốt thời gian qua.

Những kiến thức và kinh nghiệm quý báu mà chúng em có được hôm nay chính là kết quả của sự nỗ lực không ngừng của bản thân, cùng với sự tận tâm, cống hiến của Quý Thầy/Cô.

Chúng em xin kính chúc Quý Thầy/Cô dồi dào sức khỏe, hạnh phúc và luôn đạt được nhiều thành công trong sự nghiệp trồng người cao quý.

Hà Nội, ngày tháng năm 2025

Sinh viên thực hiện Sinh viên thực hiện

**Hà Gia Kiệt Trịnh Đức Kiên**

MỤC LỤC

*Trang*

[LỜI CAM ĐOAN i](#_Toc194266505)

[LỜI CẢM ƠN ii](#_Toc194266506)

[MỤC LỤC iii](#_Toc194266507)

[DANH SÁCH CÁC TỪ VIẾT TẮT vii](#_Toc194266508)

[DANH MỤC CÁC BẢNG viii](#_Toc194266509)

[DANH MỤC CÁC HÌNH ix](#_Toc194266510)

[PHẦN MỞ ĐẦU 1](#_Toc194266511)

[CHƯƠNG I: TỔNG QUAN VỀ NGÔN NGỮ TỰ NHIÊN 3](#_Toc194266512)

[1.1. Tổng quan 3](#_Toc194266513)

[1.2. Cơ sở khoa học 4](#_Toc194266514)

[1.2.1. Một số khái niệm cơ bản 4](#_Toc194266515)

[1.2.1.1. Ngôn ngữ tự nhiên 4](#_Toc194266516)

[1.2.1.2. Xử lý ngôn ngữ tự nhiên 4](#_Toc194266517)

[1.2.1.3. Trí tuệ nhân tạo 4](#_Toc194266518)

[1.2.1.4. Nhập nhằng 5](#_Toc194266519)

[1.2.1.5. Dịch máy 5](#_Toc194266520)

[1.2.2. Xác xuất (Probability) 5](#_Toc194266521)

[1.2.2.1. Thực nghiệm và không gian mẫu 5](#_Toc194266522)

[1.2.2.2. Sự kiện (Events) 5](#_Toc194266523)

[1.2.2.3. Xác suất (Probability) 6](#_Toc194266524)

[1.2.2.4. Ước lượng Xác suất 6](#_Toc194266525)

[1.2.2.5. Kỳ vọng (Expectation) và Phương sai (Variance) 6](#_Toc194266526)

[1.2.3. Lý thuyết thông tin(Information Theory) 6](#_Toc194266527)

[1.2.3.1. Khái niệm 6](#_Toc194266528)

[1.2.3.2. Entropy 6](#_Toc194266529)

[1.2.3.3. Perplexity - Cross Entropy 6](#_Toc194266530)

[1.3. Sự tiến hóa của ngôn ngữ 8](#_Toc194266531)

[1.4. Cơ sở của ngôn ngữ 8](#_Toc194266532)

[1.5. Giao tiếp và sử dụng ngôn ngữ hình thức 9](#_Toc194266533)

[1.6. Xử lý và hiểu văn bản 10](#_Toc194266534)

[1.6.1. Truy nhập cơ sở dữ liệu 10](#_Toc194266535)

[1.6.2. Thu thập thông tin 11](#_Toc194266536)

[1.6.3. Xử lý văn bản 12](#_Toc194266537)

[1.6.4. Lấy dữ liệu vào văn bản 12](#_Toc194266538)

[1.7. Các ứng dụng của ngôn ngữ tự nhiên 15](#_Toc194266539)

[1.7.1. Nhận dạng chữ viết 15](#_Toc194266540)

[1.7.2. Nhận dạng giọng nói 15](#_Toc194266541)

[1.7.3. Tổng hợp tiếng nói 15](#_Toc194266542)

[1.7.4. Dịch tự động 15](#_Toc194266543)

[1.7.5. Tìm kiếm thông tin 16](#_Toc194266544)

[1.7.6. Tóm tắt văn bản tự động 16](#_Toc194266545)

[1.7.7. Khai phá dữ liệu 16](#_Toc194266546)

[1.7.8. Truy xuất và trích chọn thông tin 16](#_Toc194266547)

[1.7.9. Trả lời câu hỏi 17](#_Toc194266548)

[1.7.10. Dịch máy 17](#_Toc194266549)

[CHƯƠNG II: XỬ LÝ NGÔN NGỮ TỰ NHIÊN 18](#_Toc194266550)

[2.1. Thách thức cơ bản trong xử lý ngôn ngữ tự nhiên 18](#_Toc194266551)

[2.1.1. Phân tích hình vị 18](#_Toc194266552)

[2.1.2. Phân tích cú pháp 18](#_Toc194266553)

[2.1.3. Phân tích ngữ nghĩa 18](#_Toc194266554)

[2.1.4. Phân tích diễn ngôn 19](#_Toc194266555)

[2.2. Qui trình xử lý ngôn ngữ tự nhiên 19](#_Toc194266556)

[2.2.1. Phân tích từ vựng (Lexical Analysis) 19](#_Toc194266557)

[2.2.2. Phân tích cú pháp (Syntax Analysis) 20](#_Toc194266558)

[2.2.3. Phân tích ngữ nghĩa (Semantic Analysis) 21](#_Toc194266559)

[2.2.4. Các giai đoạn của trình biên dịch 21](#_Toc194266560)

[2.2.5. Một số phương pháp phân tích cú pháp 23](#_Toc194266561)

[2.2.5.1. Top-down 23](#_Toc194266562)

[2.2.5.2. Bottom-up 23](#_Toc194266563)

[2.2.5.3. CYK (Cocke-Younger-Kasami) 23](#_Toc194266564)

[2.3. Các thì trong tiếng Anh 26](#_Toc194266565)

[2.4. Cách sử dụng một số thì 26](#_Toc194266566)

[2.4.1. Thì hiện tại đơn (The Simple Present Tense) 26](#_Toc194266567)

[2.4.1.1. Hình thức 26](#_Toc194266568)

[2.4.1.2. Cách sử dụng 27](#_Toc194266569)

[2.4.2. Thì hiện tại tiếp diễn (The present continuous/progressive tense) 27](#_Toc194266570)

[2.4.2.1. Hình thức 27](#_Toc194266571)

[2.4.2.2. Cách sử dụng 28](#_Toc194266572)

[2.4.3. Thì hiện tại hoàn thành (The Present Perfect Tense) 28](#_Toc194266573)

[2.4.3.1. Hình thức 28](#_Toc194266574)

[2.4.3.2. Cách sử dụng 29](#_Toc194266575)

[2.4.4. Thì hiện tại hoàn thành tiếp diễn (The Present Perfect continuousTense) 29](#_Toc194266576)

[2.4.4.1. Hình thức 29](#_Toc194266577)

[2.4.4.2. Cách sử dụng 30](#_Toc194266578)

[2.4.5. Thì quá khứ đơn (The Simple Past Tense) 30](#_Toc194266579)

[2.4.5.1. Hình thức 30](#_Toc194266580)

[2.4.5.2. Cách sử dụng 31](#_Toc194266581)

[2.4.6. Thì quá khứ tiếp diễn (The Past continuous Tense) 31](#_Toc194266582)

[2.4.6.1. Hình thức 31](#_Toc194266583)

[2.4.6.2. Cách sử dụng 31](#_Toc194266584)

[2.4.7. Thì quá khứ hoàn thành (The Past Perfect Tense) 32](#_Toc194266585)

[2.4.7.1. Hình thức 32](#_Toc194266586)

[2.4.7.2. Cách sử dụng 32](#_Toc194266587)

[2.4.8. Thì quá khứ hoàn thành tiếp diễn (The Past Perfect Continuous Tense) 33](#_Toc194266588)

[2.4.8.1. Hình thức 33](#_Toc194266589)

[2.4.8.2. Cách sử dụng 33](#_Toc194266590)

[2.4.9. Thì tương lai đơn (The Simple Future Tense) 33](#_Toc194266591)

[2.4.9.1. Hình thức 33](#_Toc194266592)

[2.4.9.2. Cách sử dụng 34](#_Toc194266593)

[2.4.10. Thì tương lai tiếp diễn (The Continuous Future Tense) 34](#_Toc194266594)

[2.4.10.1. Hình thức 34](#_Toc194266595)

[2.4.10.2. Cách sử dụng 35](#_Toc194266596)

[2.4.11. Thì tương lai hoàn thành (The Perfect Future Tense) 35](#_Toc194266597)

[2.4.11.1. Hình thức 35](#_Toc194266598)

[2.4.11.2. Cách sử dụng 36](#_Toc194266599)

[2.4.12. Thì tương lai hoàn thành tiếp diễn (The Perfect Continuous Future Tense) 36](#_Toc194266600)

[2.4.12.1. Hình thức 36](#_Toc194266601)

[2.4.12.2. Cách sử dụng 36](#_Toc194266602)

[2.5. Các cấu trúc trong câu 37](#_Toc194266603)

[2.5.1. Cấu trúc chung của 1 câu trong Tiếng Anh 37](#_Toc194266604)

[2.5.2. Tổng hợp 100 cấu trúc câu Tiếng Anh 39](#_Toc194266605)

[CHƯƠNG III: CHƯƠNG TRÌNH THỰC NGHIỆM 49](#_Toc194266606)

[3.1. 49](#_Toc194266607)

[3.2. 49](#_Toc194266608)

[3.2.1. 49](#_Toc194266609)

[3.2.1.1. 49](#_Toc194266610)

[KẾT LUẬN VÀ KIẾN NGHỊ 50](#_Toc194266611)

[1. Kết luận 50](#_Toc194266612)

[1. Kiến nghị 51](#_Toc194266613)

[TÀI LIỆU THAM KHẢO 52](#_Toc194266614)

DANH SÁCH CÁC TỪ VIẾT TẮT

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Từ viết tắt** | **Tên đầy đủ** | **Ý nghĩa** |
| 1 | KTCK | Kỹ thuật cơ khí | Kỹ thuật cơ khí |
| 2 | KCN | Khoa công nghệ | Khoa công nghệ |
| 3 | CYK | Cocke – Younger – Kasami | Thuật toán Cocke – Younger – Kasami |
| 4 | CNF | Chomsky Normal Form | Dạng chuẩn Chomsky |
| 5 | CFG | Context-Free Grammar | Văn phạm phi ngữ cảnh |
| 6 | NLP | Natural Language Processing | Xử lý ngôn ngữ tự nhiên |
| 7 | NLU | Natural Language Understanding | Hiểu ngôn ngữ tự nhiên |
| 8 | NLG | Natural Language Generation | Sinh ngôn ngữ tự nhiên |
| 9 | RBMT | Rule-Based Machine Translation | Phương pháp dựa trên quy tắc |
| 10 | EBMT | Example-Based Machine Translation | Phương pháp dựa trên ví dụ |
| 11 | BNF | British National Formulary | Ký pháp |
| 12 | IR | Information Retrieval | Truy xuất thông tin |
| 13 | IE | Informaiton Extration | Trích chọn thông tin |
| 14 | MT | Machine Translation | Dịch máy |
| 15 | CCG | Combinatory Categorial Grammar | Văn phạm danh mục kết nối |
| 16 | DG | Dependency Grammar | Văn phạm phụ thuộc |
| 17 | SMT | Statistical Translation Method | Phương pháp thống kê |
| 18 | AI | Artificial Intelligence | Trí tuệ nhân tạo |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

DANH MỤC CÁC BẢNG

*Trang*

[Bảng 2.1: Các thì trong tiếng Anh 26](#_Toc194266435)

[Bảng 2.2: Cách sử dụng thì Hiện tại đơn 27](#_Toc194266436)

[Bảng 2.3: Cách sử dụng thì Hiện tại tiếp diễn 28](#_Toc194266437)

[Bảng 2.4: Cách sử dụng thì Hiện tại hoàn thành 28](#_Toc194266438)

[Bảng 2.5: Cách sử dụng thì Hiện tại hoàn thành tiếp diễn 29](#_Toc194266439)

[Bảng 2.6: Cách sử dụng thì Quá khứ đơn 30](#_Toc194266440)

[Bảng 2.7: Cách sử dụng thì Quá khứ tiếp diễn 31](#_Toc194266441)

[Bảng 2.8: Cách sử dụng thì Quá khứ hoàn thành 32](#_Toc194266442)

[Bảng 2.9: Cách sử dụng thì Quá khứ tiếp diễn 33](#_Toc194266443)

[Bảng 2.10: Cách sử dụng thì Tương lai đơn 33](#_Toc194266444)

[Bảng 2.11: Cách sử dụng thì Tương lai tiếp diễn 35](#_Toc194266445)

[Bảng 2.12: Cách sử dụng thì Tương lai hoàn thành 35](#_Toc194266446)

[Bảng 2.12: Cách sử dụng thì Tương lai hoàn thành tiếp diễn 36](#_Toc194266447)

[Bảng 2.13: Cấu trúc thường gặp của một câu Tiếng Anh 37](#_Toc194266448)

[Bảng 3.1: Nội dung bảng tô đen thứ tự, dùng style “Bang”. 49](#_Toc194266449)

DANH MỤC CÁC HÌNH

*Trang*

[Hình 1.1: Tín hiệu sóng âm của một câu tiếng Việt 3](#_Toc194266476)

[Hình 1.2: Hai đối tượng giao tiếp với ngôn ngữ 10](#_Toc194266477)

[Hình 1.3: Sơ đồ phân tích câu “The agent feels a breeze” 13](#_Toc194266478)

[Hình 2.1: Một trình biên dịch 19](#_Toc194266479)

[Hình 2.2: Một cây phân tích cú pháp 20](#_Toc194266480)

[Hình 2.3: Chuyển đổi kiểu trên cây phân tích cú pháp 21](#_Toc194266481)

[Hình 2.4: Các giai đoạn của một trình biên dịch 22](#_Toc194266482)

[Hình 2.5: Phân tích ví dụ CYK 24](#_Toc194266483)

[Hình 2.6: Ngữ pháp tiếng Anh thu nhỏ 24](#_Toc194266484)

[Hình 2.7: Ngữ pháp và cách chuyển đổi sang CNF 25](#_Toc194266485)

[Hình 2.8: Thử sinh ra một văn phạm tương ứng 25](#_Toc194266486)

[Hình 2.9: Thuật toán parsing CYK 25](#_Toc194266487)

[Hình 3.1: Nội dung hình tô đen thứ tự, dùng style “Hinh” 49](#_Toc194266488)

PHẦN Mở đầu

***Tầm quan trọng ý nghĩa của đề tài, lý do chọn đề tài***

Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) là một lĩnh vực then chốt của trí tuệ nhân tạo, đóng vai trò cầu nối giữa máy tính và ngôn ngữ con người. Trong bối cảnh chuyển đổi số và phát triển công nghệ thông tin mạnh mẽ hiện nay, NLP ngày càng khẳng định vai trò quan trọng của mình trong việc giải quyết nhiều bài toán thực tiễn.

Tầm quan trọng của đề tài này thể hiện qua khả năng ứng dụng rộng rãi trong nhiều lĩnh vực như dịch thuật tự động, hệ thống trợ lý ảo, phân tích cảm xúc người dùng, tóm tắt văn bản và nhiều ứng dụng khác. Đặc biệt, trong thời đại bùng nổ thông tin, NLP giúp tự động hóa việc xử lý khối lượng dữ liệu ngôn ngữ khổng lồ, tiết kiệm thời gian và nguồn lực con người.

Lý do chọn đề tài xuất phát từ những thách thức đặc thù khi xử lý ngôn ngữ tự nhiên - loại dữ liệu không tuân theo cấu trúc cố định, đòi hỏi các phương pháp và thuật toán chuyên biệt. Việc nghiên cứu sâu về NLP không chỉ mang lại giá trị học thuật mà còn mở ra nhiều cơ hội phát triển ứng dụng thực tiễn, đáp ứng nhu cầu ngày càng cao của xã hội trong việc tương tác với máy tính thông qua ngôn ngữ tự nhiên.

***Mục tiêu nghiên cứu***

Đề tài đặt ra các mục tiêu cụ thể sau:

Nghiên cứu tổng quan về các phương pháp và kỹ thuật hiện đại trong xử lý ngôn ngữ tự nhiên

Phân tích các thách thức trong việc xử lý cấu trúc và ngữ pháp của dữ liệu ngôn ngữ đầu vào

Xây dựng mô hình phát hiện và sửa lỗi ngôn ngữ tự động

Đánh giá hiệu quả của các phương pháp NLP trong các ứng dụng thực tế như dịch máy và phân tích văn bản

***Đối tượng và phạm vi nghiên cứu của đề tài***

Đối tượng nghiên cứu:

Các kỹ thuật và thuật toán xử lý ngôn ngữ tự nhiên

Dữ liệu ngôn ngữ dưới dạng văn bản và giọng nói

Các mô hình hiểu và phân tích ngữ nghĩa, ngữ cảnh trong ngôn ngữ tự nhiên

Phạm vi nghiên cứu:

Tập trung vào các phương pháp xử lý ngôn ngữ tự nhiên dựa trên học máy và học sâu

Giới hạn nghiên cứu trong phạm vi xử lý văn bản, không mở rộng sang nhận dạng giọng nói

Thực nghiệm trên bộ dữ liệu tiếng Việt và tiếng Anh

Đánh giá hiệu quả của các mô hình trong khoảng thời gian 6 tháng

***Phương pháp nghiên cứu***

Để thực hiện đề tài, các phương pháp nghiên cứu sau sẽ được áp dụng:

*Phương pháp nghiên cứu tài liệu:* Thu thập và phân tích các công trình nghiên cứu, bài báo khoa học và tài liệu liên quan đến xử lý ngôn ngữ tự nhiên để xây dựng nền tảng lý thuyết vững chắc.

*Phương pháp mô hình hóa:* Xây dựng các mô hình toán học và thuật toán để biểu diễn và xử lý dữ liệu ngôn ngữ tự nhiên.

*Phương pháp thực nghiệm:* Triển khai và đánh giá các mô hình NLP trên các bộ dữ liệu thực tế, so sánh hiệu quả của các phương pháp khác nhau.

*Phương pháp phân tích định lượng:* Sử dụng các chỉ số đo lường như độ chính xác, độ bao phủ, F1-score để đánh giá khách quan hiệu quả của các mô hình.

*Phương pháp so sánh đối chiếu:* So sánh kết quả nghiên cứu với các phương pháp hiện có để xác định ưu điểm và hạn chế của từng phương pháp.

CHƯƠNG I: TỔNG QUAN VỀ NGÔN NGỮ TỰ NHIÊN

1.1. Tổng quan

Xử lý ngôn ngữ tự nhiên là một lĩnh vực chuyên về xử lý thông tin có đầu vào là "dữ liệu ngôn ngữ", bao gồm cả "văn bản" và "tiếng nói". Đây là hai dạng dữ liệu đang dần trở thành nguồn thông tin chủ yếu mà con người tạo ra và lưu trữ dưới dạng điện tử. Đặc trưng nổi bật của chúng là không có cấu trúc hoặc chỉ có cấu trúc một phần, nên không thể tổ chức trong các định dạng cố định như bảng biểu truyền thống. Theo đánh giá từ Oracle, khoảng 80% dữ liệu toàn cầu hiện nay là dữ liệu không cấu trúc [Oracle Text].

Với sự phát triển của Internet, sách báo điện tử, máy tính cá nhân, thiết bị viễn thông, và công cụ ghi âm, bất kỳ ai cũng có thể tạo ra văn bản hoặc tín hiệu âm thanh. Thách thức đặt ra là làm sao xử lý và biến đổi chúng từ dạng "chưa hiểu được" thành dạng "có thể hiểu và khai thác", từ đó trích xuất thông tin và tri thức có giá trị.

Ví dụ, khi ta có những câu sau bằng các ngôn ngữ khác nhau:

* "We meet here today to talk about Vietnamese language and speech processing."
* "Aujourd'hui nous nous réunissons ici pour discuter le traitement de langue et de parole vietnamienne."
* "Мы встречаемся здесь сегодня, чтобы говорить о вьетнамском языке и обработке речи."

Nếu có người hoặc một chương trình máy tính dịch các câu này sang tiếng Việt, ta sẽ hiểu ý nghĩa của chúng là: "Hôm nay chúng ta gặp nhau ở đây để bàn về xử lý ngôn ngữ và tiếng nói tiếng Việt." Những câu này, khi được lưu dưới dạng tệp văn bản, sẽ là dữ liệu "văn bản". Ngược lại, nếu chúng được đọc lên và ghi âm lại, chúng sẽ trở thành các tệp "tiếng nói" dưới dạng tín hiệu âm thanh. Tín hiệu sóng âm của 1 câu tiếng Việt có thể nhìn thấy như sau:



Hình 1.1: Tín hiệu sóng âm của một câu tiếng Việt

Tuy nhiên, thực tế, một bài báo khoa học có thể chứa hàng nghìn câu, và trên thế giới có hàng triệu văn bản. Web, cùng với các thư viện điện tử, là nguồn dữ liệu văn bản khổng lồ. Trong tương lai gần, khi toàn bộ sách báo, âm thanh cũ được số hóa (nhờ công nghệ nhận dạng chữ viết, thu âm, hoặc nhập liệu), máy tính sẽ trở thành kho lưu trữ gần như toàn bộ tri thức nhân loại. Thách thức quan trọng là làm thế nào để xử lý và chuyển đổi khối lượng dữ liệu văn bản và tiếng nói khổng lồ này, giúp mọi người tiếp cận thông tin và tri thức cần thiết.

Hiện nay, xử lý ngôn ngữ tự nhiên đã được ứng dụng vào nhiều lĩnh vực thực tiễn như: nhận dạng chữ viết, nhận dạng tiếng nói, tổng hợp giọng nói, dịch tự động, tìm kiếm thông tin, tóm tắt văn bản, khai phá dữ liệu, và phát hiện tri thức.

1.2. Cơ sở khoa học

1.2.1. Một số khái niệm cơ bản

1.2.1.1. Ngôn ngữ tự nhiên

Ngôn ngữ là hệ thống để giao thiệp hay suy luận dùng một cách biểu diễn phép ẩn dụ và một loại ngữ pháp theo logic, mỗi cái đó bao hàm một tiêu chuẩn hay sự thật thuộc lịch sử và siêu việt. Nhiều ngôn ngữ sử dụng điệu bộ, âm thanh, ký hiệu, hay chữ viết, và cố gắng truyền khái niệm, ý nghĩa, và ý nghĩ, nhưng mà nhiều khi những khía cạnh này nằm sát quá, cho nên khó phân biệt nó.

1.2.1.2. Xử lý ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên (natural language processing - NLP) là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ - công cụ hoàn hảo nhất của tư duy và giao tiếp.

1.2.1.3. Trí tuệ nhân tạo

Trí tuệ nhân tạo hay trí thông minh nhân tạo (tiếng Anh: artificial intelligence hay machine intelligence, thường được viết tắt là AI) là trí tuệ được biểu diễn bởi bất cứ một hệ thống nhân tạo nào. Thuật ngữ này thường dùng để nói đến các máy tính có mục đích không nhất định và ngành khoa học nghiên cứu về các lý thuyết và ứng dụng của trí tuệ nhân tạo.

1.2.1.4. Nhập nhằng

Nhập nhằng (ambiguity) là hiện tượng một từ, cụm từ hoặc câu có thể được hiểu theo nhiều cách khác nhau, là một trong những thách thức lớn nhất trong xử lý ngôn ngữ tự nhiên. Hiện tượng này xuất hiện ở nhiều cấp độ của ngôn ngữ, bao gồm: nhập nhằng từ vựng khi một từ có nhiều nghĩa khác nhau, nhập nhằng cú pháp khi một câu có thể được phân tích theo nhiều cấu trúc ngữ pháp khác nhau, nhập nhằng ngữ nghĩa khi ý nghĩa của câu có thể được hiểu theo nhiều cách khác nhau, và nhập nhằng ngữ dụng khi ý nghĩa phụ thuộc vào ngữ cảnh sử dụng. Việc giải quyết vấn đề nhập nhằng đòi hỏi các hệ thống xử lý ngôn ngữ tự nhiên phải có khả năng phân tích ngữ cảnh và tích hợp nhiều nguồn thông tin khác nhau để đưa ra giải thích phù hợp nhất.

1.2.1.5. Dịch máy

Dịch máy là công nghệ cho phép chuyển đổi tự động nội dung từ một ngôn ngữ sang ngôn ngữ khác, có thể là một phần hoặc toàn bộ văn bản. Trong lĩnh vực này, có ba phương pháp chính được sử dụng rộng rãi: phương pháp dựa trên ví dụ (EBMT), phương pháp dựa trên luật (RBMT), và phương pháp thống kê (SMT). Trong những năm gần đây, phương pháp thống kê đã trở thành trọng tâm nghiên cứu nhờ những ưu điểm vượt trội so với các phương pháp khác. Trong số các mô hình dịch máy thống kê, bốn loại đã chứng minh được hiệu quả vượt trội: mô hình dịch theo từ, mô hình dịch theo cú pháp, mô hình dịch theo cụm từ, và mô hình dịch theo cụm từ phân cấp.

1.2.2. Xác xuất (Probability)

1.2.2.1. Thực nghiệm và không gian mẫu

Không gian mẫu (ký hiệu Ω) là tập hợp tất cả các kết quả có thể có của một thực nghiệm. Có nhiều ví dụ về không gian mẫu như: tung đồng xu Ω = {head, tail}, bầu cử Ω = {yes/no}, tung xúc xắc Ω = {1,...,6}, xổ số (|Ω| ≈ 107 .. 1012), số lượng tai nạn giao thông/năm (Ω = N), hay trong lỗi chính tả (Ω = Z\*, với Z là bảng chữ cái).

1.2.2.2. Sự kiện (Events)

Sự kiện A là một tập các mẫu A ⊆ Ω, và tập tất cả A là 2Ω. Trong đó Ω là sự kiện chắc chắn, Ø là sự kiện không xảy ra. Ví dụ khi tung đồng xu 3 lần, Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}, và sự kiện có đúng 2 lần xuất hiện Tail là A = {HTT, THT, TTH}.

1.2.2.3. Xác suất (Probability)

Xác suất được xác định thông qua việc thực hiện một thực nghiệm nhiều lần và đếm số lần sự kiện A xảy ra ("count" c1). Mỗi lần thực nghiệm được gọi là một dãy. Khi thực hiện nhiều dãy và ghi nhận các giá trị ci, tỉ số ci/Ti (Ti là tổng số lần thực nghiệm trong dãy thứ i) sẽ dần hội tụ về một hằng số. Giá trị hằng số này được gọi là xác suất của A, ký hiệu p(A).

1.2.2.4. Ước lượng Xác suất

Có hai cách ước lượng xác suất:

* Từ một dãy thực nghiệm: p(A) = c1/T1
* Nếu thực hiện được nhiều dãy thực nghiệm: tính trung bình cộng của ci/Ti

1.2.2.5. Kỳ vọng (Expectation) và Phương sai (Variance)

Kỳ vọng là tổng trọng số của giá trị của X, hay giá trị trung bình của biến ngẫu nhiên, được tính bằng: .

Phương sai là trung bình bình phương của độ lệch (độ lệch của biến X so với trung bình của nó), được tính bằng: .

1.2.3. Lý thuyết thông tin(Information Theory)

1.2.3.1. Khái niệm

Lý thuyết thông tin nghiên cứu về việc áp dụng các công cụ toán học trong việc lượng hóa data cho mục đích lưu trữ và truyền dữ liệu. Độ đo thông tin là Entropy, là số lượng bít trung bình cần thiết cho việc lưu trữ hay truyền dữ liệu. Nó đóng vai trò quan trọng trong xử lý thông tin bằng các phương pháp thống kê, đặc biệt trong NLP.

1.2.3.2. Entropy

Entropy là một độ đo thông tin. Entropy ~ hỗn độn, mờ, trái nghĩa với order, … Đo độ không chắc chắn: Entropy thấp 🡪 Đo độ không chắc chắn thấp; Entropy cao 🡪 Đo độ không chắc chắn cao. Trong vật lý: Entropy giảm khi năng lượng được sử dụng. Ký hiệu p(x) là một phân bố của một biến ngẫu nhiên X. Ω là không gian mẫu của X. Entropy được tính như sau: . Đơn vị: bits (log10: nats). Kí hiệu: .

1.2.3.3. Perplexity - Cross Entropy

*a. Entropy liên quan thế nào đến hiểu ngôn ngữ?*

Trong lĩnh vực xử lý ngôn ngữ tự nhiên, mối quan hệ giữa entropy và hiểu ngôn ngữ được thể hiện thông qua mức độ chính xác của thông tin. Khi một vấn đề chứa nhiều thông tin xác thực, entropy sẽ có giá trị thấp. Entropy cũng được sử dụng để đánh giá hiệu quả của các mô hình ngôn ngữ khác nhau. Ví dụ, trong một mô hình mã hóa ký tự cơ bản (0-gram), số bit trung bình cần thiết cho mỗi ký tự là 2.5, nhưng khi xét đến mối liên hệ giữa các âm tiết, ta có thể tạo ra mô hình tốt hơn với entropy chỉ còn 1.22 bit trên ký tự.

*b. Perplexity*

Được định nghĩa là 2H với H là entropy của phân bố , đại diện cho số lượng lựa chọn trung bình mà một biến cần thực hiện. Giá trị perplexity càng nhỏ (tương ứng với entropy thấp) thể hiện mô hình càng hiệu quả, nghĩa là cần ít bit hơn để mã hóa thông tin. Ta có thể minh họa điều này qua ví dụ:

Xác suất lựa chọn 8 con ngựa đua trong đó:

Ngựa 1 có xác suất: 1/2, ngựa 2: 1/4, ngựa 3: 1/8, ngựa 4: 1/16, và bốn ngựa còn lại mỗi con có xác suất: 1/64.

*c. Entropy rate*

Tính entropy của một dãy các từ trong một ngôn ngữ L.

Entropy rate được coi như per-word entropy.Coi một ngôn ngữ như một quá trình ngẫu nhiên sản xuất một dãy các từ. Cần quan tâm đến một dãy vô hạn từ. Entropy rate H(L) được định nghĩa như sau:

*d. Cross Entropy*

Cross entropy được sử dụng khi chúng ta không biết phân bố thật p.

Cross-entropy của phân bố m của phân bố thật p được định nghĩa:

(Theo lý thuyết Shannon-McMillan-Breiman)

*e. Cross entropy để so sánh các mô hình :*

Cross entropy là cận trên của entropy Mô hình m càng chính xác thì cross entropy càng gần với entropy Độ khác nhau và đo độ chính xác của mô hình m.

*f. Các công thức Cross Entropy*

Cross entropy giữa biến X với phân bố xác suất đúng p(x) và một phân bố m được tính như sau:

1.3. Sự tiến hóa của ngôn ngữ

Để hiểu lời nói hành động, chúng ta cần tiếp cận vấn đề tương tự như cách hiểu hình ảnh hay chẩn đoán y học. Quá trình này bắt đầu bằng việc tiếp nhận một tập hợp các đầu vào đa nghĩa, từ đó chúng ta suy luận để xác định trạng thái nào của thế giới có thể được phản ánh qua những đầu vào đó.

Hiểu lời nói hành động là một phần quan trọng trong việc đặc tả ngôn ngữ. Một khía cạnh của việc này có thể được giải thích thông qua lý do logic. Chúng ta nhận thấy rằng các chuỗi logic liên kết với nhau là cách hiệu quả để mô tả cách mà từ ngữ và cụm từ kết hợp để tạo thành một cụm từ lớn hơn.

Tuy nhiên, một phần khác của việc hiểu vấn đề có thể chỉ được giải thích bằng các lý do kỹ thuật không rõ ràng. Thực tế, có nhiều trạng thái của thế giới có thể dẫn đến một lời nói hành động tương tự, do đó, người hiểu cần phải xác định trạng thái nào là khả thi hơn.

1.4. Cơ sở của ngôn ngữ

Bản chất của ngôn ngữ hình thức có thể được hiểu như một tập hợp các chuỗi ký tự, trong đó mỗi chuỗi được tạo thành từ các ký hiệu cơ bản (terminal symbols) thuộc một tập hữu hạn. Mặc dù có nhiều cách tiếp cận và ký hiệu khác nhau trong việc mô tả ngữ pháp của cả ngôn ngữ tự nhiên và ngôn ngữ hình thức, phần lớn đều dựa trên nguyên tắc cấu trúc cụm từ - nơi các chuỗi ký tự được tổ chức thành các đơn vị lớn hơn gọi là cụm từ, tạo nên các phạm trù khác nhau.

Trong đó, các phạm trù như cụm danh từ (NP) và cụm động từ (VP) được gọi là các ký hiệu không kết thúc (nonterminal symbols). Khi sử dụng ký pháp BNF, quy tắc viết được tổ chức với ký hiệu không kết thúc ở vế trái và các thành phần liên kết ở vế phải. Ví dụ về một quy tắc điển hình: S → NP VP, nghĩa là một câu có thể được tạo thành bằng cách ghép một cụm danh từ với một cụm động từ.

Về quy trình giao tiếp, có thể phân tích thành 7 bước chính trong tương tác giữa người nói (S) và người nghe (H):

***Đối với người nói:***

Mục đích: S mong muốn H tin vào thông điệp P (với điều kiện S đã tin P).

Phát sinh: S lựa chọn từ ngữ W để truyền đạt nội dung P một cách hiệu quả.

Tổng hợp: S phát âm các từ W để truyền đạt trực tiếp đến H.

***Đối với người nghe:***

Nhận thức: H tiếp nhận thông điệp W' (lý tưởng là W' = W, nhưng có thể có sai lệch)

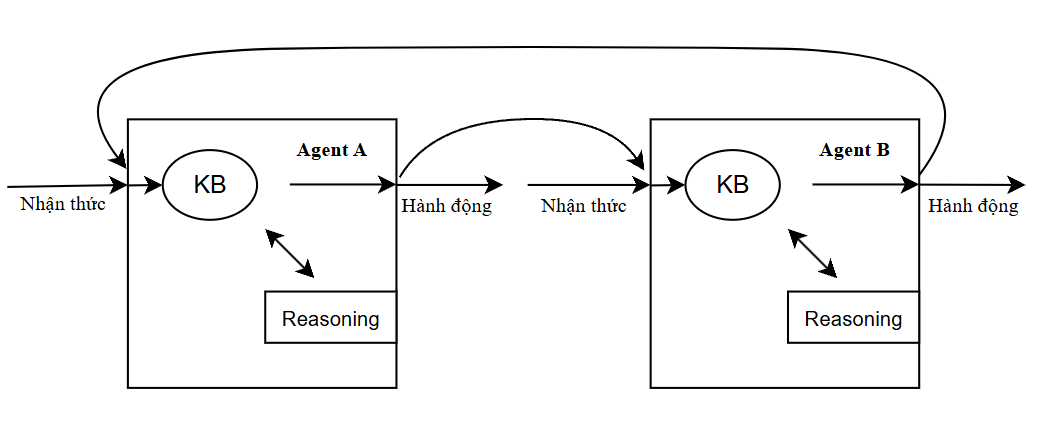
Phân tích: H xử lý W' để hiểu các nghĩa có thể P1, P2,... (do từ và cụm từ có thể đa nghĩa).

Ý định: H nhận định rằng S muốn truyền đạt P (lý tưởng là P = P ban đầu, nhưng có thể có sự khác biệt trong diễn giải).

Hợp nhất: H quyết định chấp nhận hay từ chối thông điệp Pi dựa trên mức độ tin cậy

1.5. Giao tiếp và sử dụng ngôn ngữ hình thức

Trong quá trình tương tác, các thực thể thường sử dụng ngôn ngữ làm phương tiện giao tiếp thay vì trao đổi trực tiếp thông tin nền tảng. Quá trình này có thể được minh họa qua một mô hình, trong đó một thực thể tạo ra thông điệp bằng ngôn ngữ và thực thể khác có khả năng tiếp nhận. Điều đáng chú ý là ngôn ngữ được sử dụng để giao tiếp bên ngoài có thể khác biệt với cách biểu đạt nội bộ của mỗi thực thể. Không nhất thiết phải có sự đồng nhất về ký hiệu nội bộ giữa các thực thể, miễn là mỗi thực thể có thể chuyển đổi chính xác từ ngôn ngữ bên ngoài sang hệ thống ký hiệu riêng của mình.



Hình 1.2: Hai đối tượng giao tiếp với ngôn ngữ

Khi sử dụng ngôn ngữ bên ngoài để giao tiếp, hai thách thức chính xuất hiện: tạo lập và diễn giải thông điệp. Nhiều nghiên cứu về xử lý ngôn ngữ tự nhiên đã tập trung vào việc phát triển giải pháp cho hai vấn đề này. Tuy nhiên, thách thức lớn nhất nằm ở việc đảm bảo sự tương thích về kiến thức nền tảng giữa các thực thể giao tiếp. Cách một thực thể A diễn đạt và cách thực thể B hiểu phụ thuộc chủ yếu vào những hiểu biết và niềm tin của cả hai, bao gồm cả nhận thức về nhau. Ngay cả khi hai thực thể chia sẻ cùng một hệ thống ngôn ngữ, họ vẫn cần thời gian để học cách giao tiếp hiệu quả.

Sự phát triển từ lý thuyết trò chơi đến các ứng dụng ngôn ngữ thực tế đã mang lại nhiều tiến bộ. Một số vấn đề then chốt đã được giải quyết bao gồm:

Việc ứng dụng vào thực tiễn đã chứng minh hiệu quả của công nghệ xử lý ngôn ngữ tự nhiên. Các thách thức như xử lý văn bản dài, tối ưu hóa thuật toán phân tích cú pháp, mở rộng vốn từ vựng đặc biệt, xử lý cấu trúc ngữ pháp phức tạp, dịch thuật theo ngữ cảnh và hiểu đúng ngữ nghĩa đều được quan tâm giải quyết.

Các hệ thống xử lý ngôn ngữ tự nhiên thành công thường có hai đặc điểm chung: chúng tập trung vào một lĩnh vực cụ thể thay vì cố gắng bao quát tất cả, và chúng tập trung vào một nhiệm vụ nhất định thay vì cố gắng hiểu toàn bộ ngôn ngữ.

1.6. Xử lý và hiểu văn bản

1.6.1. Truy nhập cơ sở dữ liệu

Một trong những lĩnh vực đầu tiên đạt được thành công trong xử lý ngôn ngữ tự nhiên là việc truy cập vào cơ sở dữ liệu (CSDL). Vào những năm 1970, nhiều CSDL được lưu trữ trên các máy tính lớn (mainframe), nhưng việc truy cập chúng yêu cầu người dùng phải viết các chương trình phức tạp bằng những ngôn ngữ lập trình khó hiểu. Nhân viên vận hành máy không thể đáp ứng hết nhu cầu của người dùng, trong khi đó, người dùng lại không muốn học lập trình. Để giải quyết vấn đề này, giao diện ngôn ngữ tự nhiên đã được phát triển. Một trong những ví dụ đầu tiên là hệ thống LUNAR, được thiết kế bởi William Woods và nhóm của ông cho NASA vào năm 1973. Hệ thống này cho phép các nhà địa chất đặt câu hỏi về dữ liệu hóa học của các mẫu đất đá từ mặt trăng mà tàu Apollo mang về. Mặc dù không thể áp dụng trong thực tế, nhưng trong các thử nghiệm, nó đã thành công trong 78% các câu hỏi, chẳng hạn như:

" What is the average model plagioclase concentration for lunar samples that contain rubidium?"

Một hệ thống tương tự là của Fernando Pereira (1983), có khả năng trả lời các câu hỏi về CSDL địa lý như:

Q: “Which countries are bounded by two seas?”

A: “Egypt, Iran, Israel, Saudi Arabia and Turkey.”

Q: “Whats are the counties from which a river flows into Black sea?”

A: “Romania, Soviet Union.”

Mặc dù những hệ thống này mang lại nhiều lợi ích, nhưng cũng có nhược điểm là người dùng không biết khi nào câu hỏi của họ được trả lời thành công và những từ nào không nằm trong hệ thống. Vào cuối thế kỷ 20, một số hệ thống thương mại đã phát triển một kho từ vựng và ngữ pháp phong phú để đáp ứng nhu cầu của người dùng. Một thách thức lớn trong các hệ thống hiện tại là sự tương tác giữa người dùng và hệ thống, khi người dùng có thể đặt nhiều câu hỏi liên quan đến nhau.

1.6.2. Thu thập thông tin

Quá trình thu thập thông tin liên quan đến việc trích xuất dữ liệu từ văn bản để trả lời một câu hỏi cụ thể. Một số tài liệu có thể được mô tả bằng các đại diện như tiêu đề, từ khóa hoặc tóm tắt. Với lượng thông tin khổng lồ hiện có trên mạng, việc sử dụng toàn bộ văn bản và chia nó thành các đoạn nhỏ để thu thập thông tin là phương pháp hiệu quả. Các câu hỏi thường được thể hiện dưới dạng danh sách từ khóa. Trong các hệ thống thu thập thông tin ban đầu, câu hỏi thường là sự kết hợp logic của các từ khóa. Nếu một câu hỏi không tìm thấy tài liệu, có thể do nó quá hẹp. Việc thay đổi "và" thành "hoặc" hoặc thêm vào các từ khóa có thể giúp mở rộng tìm kiếm, nhưng đôi khi lại dẫn đến việc tìm thấy quá nhiều tài liệu mà không có sự hướng dẫn rõ ràng.

Hầu hết các hệ thống hiện đại đã chuyển từ phương pháp logic sang phương pháp không gian vector, trong đó danh sách từ (cả trong tài liệu và câu hỏi) được coi như một vector trong không gian n chiều, với n là số lượng đặc trưng của tập tài liệu. Việc tìm kiếm tài liệu trở thành quá trình so sánh vector này với các vector khác và xác định những vector gần nhất. Phương pháp vector linh hoạt hơn so với phương pháp logic, vì nó cho phép sắp xếp tài liệu theo khoảng cách đến câu hỏi, với tài liệu gần nhất được báo cáo trước.

Có nhiều biến thể trong phương pháp này. Một số hệ thống cho phép người dùng yêu cầu rằng hai từ phải xuất hiện gần nhau để được tính là một lần, trong khi một số khác sử dụng từ điển đồng nghĩa để mở rộng danh sách từ khóa. Chỉ những hệ thống kém nhất mới đếm tất cả các từ trong vector mà không phân biệt. Nhiều hệ thống đánh giá trọng số cho các từ khác nhau, với cách tốt nhất là tăng trọng số cho những từ đặc trưng, tức là những từ xuất hiện trong ít tài liệu hơn.

1.6.3. Xử lý văn bản

Kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) đã đạt được thành công trong việc phân loại văn bản theo các chủ đề cụ thể. Một số hệ thống thương mại đã sử dụng phương pháp này để truy cập thông tin từ các bức điện báo. Người dùng có thể yêu cầu thông tin trong các lĩnh vực như công nghiệp, thương mại hoặc địa lý. Các nhà cung cấp đã tận dụng kiến thức của các chuyên gia để xác định các lớp phân loại. Trong những năm gần đây, các hệ thống NLP đã chứng minh được độ chính xác cao, với khả năng phân loại chính xác hơn 90% thông tin thời sự. Chúng cũng nhanh chóng và hiệu quả hơn, dẫn đến sự chuyển đổi từ quy trình thủ công sang tự động hóa.

Phân loại văn bản theo các kỹ thuật NLP không giống như truy xuất thông tin (IR), vì phân loại là cố định và những người phát triển các hệ thống này đã tập trung vào việc tích hợp các chương trình của họ với vấn đề này.

1.6.4. Lấy dữ liệu vào văn bản

Quá trình trích xuất dữ liệu từ văn bản liên quan đến việc lấy ra những thông tin cần thiết để đưa vào một cấu trúc dữ liệu cụ thể.

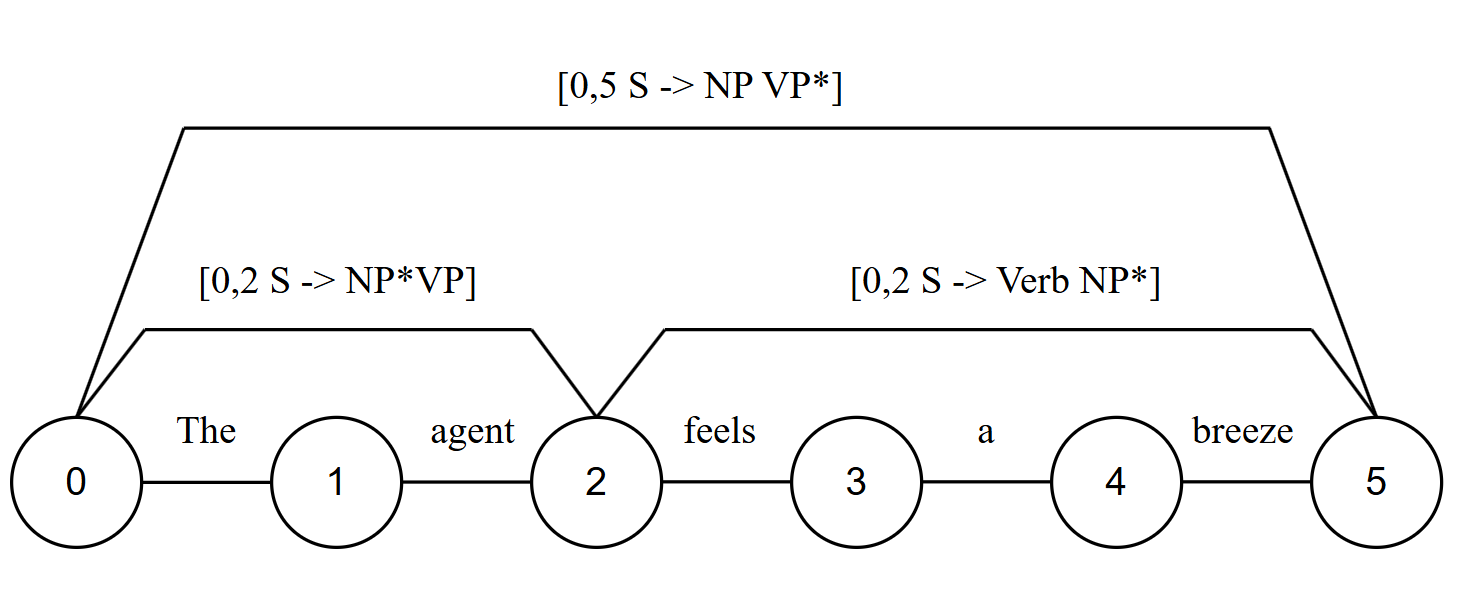
**Hiệu Quả Của Thuật Toán Phân Tích Từ**

Trong phần này, chúng ta sẽ xem xét các yếu tố ảnh hưởng đến hiệu quả của thuật toán phân tích từ. Có ba nguyên tắc chính cần lưu ý để tối ưu hóa hiệu suất:

1. Tránh thực hiện hai lần những công việc có thể hoàn thành trong một lần.
2. Không thực hiện những công việc không cần thiết.
3. Không trình bày thông tin một cách riêng lẻ nếu không cần thiết.

Chúng ta sẽ thiết kế một thuật toán phân tích từ với cách tiếp cận như sau: Ví dụ, cụm từ “the students in section 2 of Computer Science 101” được xác định là một danh từ (Noun Phrase - NP). Điều này cho thấy rằng kết quả có thể được lưu trữ trong một cấu trúc dữ liệu đã biết, gọi là sơ đồ. Các thuật toán này được gọi là phân tích từ loại theo sơ đồ. Khi làm việc với ngữ pháp ngữ cảnh tự do (context-free grammar), mọi mệnh đề tìm thấy trong một nhánh của không gian tìm kiếm cũng có thể áp dụng cho nhánh khác. Việc ghi nhận kết quả trong sơ đồ giúp lập trình viên tránh lặp lại công việc.

Chúng ta nhận thấy rằng thuật toán phân tích sơ đồ kết hợp cả hai phương pháp: xử lý từ trên xuống (top-down) và từ dưới lên (bottom-up).



Hình 1.3: Sơ đồ phân tích câu “The agent feels a breeze”

Kết quả của thuật toán là một rừng cây phân tích được đóng gói (packed forest), không chỉ đơn thuần là việc đếm mọi khả năng có thể. Sơ đồ là một cấu trúc dữ liệu mô tả các kết quả thành phần của quá trình phân tích, cho phép tái sử dụng. Một sơ đồ cho một câu có n từ sẽ có n+1 đỉnh và một số cạnh nối với các vector. Ví dụ, cạnh có nhãn [0,5, S - NP VP\*] cho thấy rằng danh từ NP (Noun Phrase) theo sau bởi động từ VP (Verb Phrase) tạo thành một mệnh đề S (Sentence) trải từ vị trí 0 đến 5. Dấu \* trong một cạnh chỉ ra rằng nó tách ra từ các phần còn lại. Các cạnh có dấu \* ở cuối được gọi là cạnh hoàn thiện; ví dụ, cạnh [0, 2 S - NP\*VP] cho biết một NP trải từ 0 đến 2, và nếu tìm thấy một VP theo sau, sẽ có một S. Ngược lại, các cạnh có dấu chấm trước dấu kết thúc được gọi là cạnh không hoàn thiện, đang tìm kiếm một VP.

Chúng ta đã biết hai cách tiếp cận trong quá trình xử lý. Trong phương pháp phân tích từ dưới lên (Bottom-Up), chúng ta mô tả quá trình xây dựng cây từ các từ, quay lui khi cần thiết. Với ngữ pháp mệnh đề nhất định (Definite Clause Grammar), chúng ta mô tả việc xử lý như một mẫu suy luận logic trong các chuỗi. Việc quay lui được sử dụng khi một số quy tắc có thể dẫn đến cùng một dự đoán. Bây giờ, chúng ta sẽ xem xét một cách tiếp cận thứ ba. Theo cách nhìn này, quá trình phân tích một câu n từ sẽ tạo ra một sơ đồ mẫu với n + 1 đỉnh và thêm vào một số cạnh để biểu diễn, cố gắng tạo ra một cạnh hoàn thiện trải từ đỉnh 0 đến đỉnh n, tạo thành một phân lớp S. Không có việc quay lui: mọi thứ đều được đặt trong sơ đồ này.

***Phân Tích Mở Rộng Từ Sơ Đồ: Đóng gói***

Khi thuật toán phân tích sơ đồ hoàn tất, nó trả về toàn bộ sơ đồ, nhưng chúng ta thực sự cần một cây (tree) hoặc một số cây. Tùy thuộc vào loại phân tích được sử dụng, chúng ta có thể muốn chọn một hoặc tất cả các cây phân tích trải toàn bộ đầu vào, hoặc chỉ xem xét một số cây con mà không cần trải ra toàn bộ đầu vào. Nếu có một ngữ pháp bổ sung, có thể chúng ta chỉ muốn tìm mở rộng ngữ nghĩa, bỏ qua các cấu trúc cú pháp. Trong mọi trường hợp, chúng ta cần khả năng phân tích mở rộng từ một sơ đồ.

Cách đơn giản nhất để thực hiện điều này là điều chỉnh bộ hoàn thiện (Completer) sao cho khi nó kết hợp hai cạnh con thành một cạnh cha. Cạnh cha sẽ chứa danh sách các cạnh con cấu thành nó. Sau đó, trong quá trình phân tích, chỉ cần tìm trong chart[n] cho một cạnh bắt đầu tại 0, đệ quy qua danh sách các cạnh con để tạo ra cây phân tích hoàn chỉnh. Quyết định về việc thực hiện phân tích mở rộng sẽ phụ thuộc vào các quy tắc đã được xác định.

Chúng ta kết thúc phần này bằng việc phân tích độ phức tạp của thuật toán, với độ phức tạp O(n³) trong trường hợp xấu nhất (với n là số từ đầu vào). Trường hợp tốt nhất có thể đạt được với ngữ pháp ngữ cảnh tự do. Lưu ý rằng nếu không có rừng đóng gói, thuật toán có thể gặp vấn đề trong trường hợp xấu nhất, vì có thể có O(2ⁿ) cây phân tích khác nhau. Trong thực tế, thuật toán có thể thực hiện phân tích với tốc độ 100 từ mỗi giây, tùy thuộc vào độ phức tạp của ngữ pháp và đầu vào.

***Dấu Hiệu Cú Pháp***

Sự thay đổi trong các từ như động từ và giới từ có thể tạo ra nhiều sự nhập nhằng, vì chúng có thể dẫn đến những ý nghĩa khác nhau. Ví dụ: "Lee asked Kim to tell Toby to leave on Saturday." Ở đây, phó từ "Saturday" có thể được liên kết với cả "tell" và "leave".

***Dấu Hiệu Từ Vựng***

Có nhiều từ có thể gây nhầm lẫn, nhưng không phải tất cả đều giống nhau. Khi hỏi về nghĩa của từ "pen", hầu hết mọi người sẽ trả lời rằng đó là một công cụ để viết. Tuy nhiên, từ này còn có những nghĩa khác như hàng rào, nhà giam, hoặc con thiên nga đực.

1.7. Các ứng dụng của ngôn ngữ tự nhiên

1.7.1. Nhận dạng chữ viết

Có hai loại nhận diện chữ viết: đầu tiên là nhận diện chữ in, chẳng hạn như việc chuyển đổi văn bản từ sách giáo khoa thành định dạng điện tử, ví dụ như file doc của Microsoft Word. Loại thứ hai phức tạp hơn là nhận diện chữ viết tay, do chữ viết tay không có hình thức đồng nhất và có sự khác biệt giữa các cá nhân. Các chương trình nhận diện chữ in có khả năng chuyển đổi hàng ngàn cuốn sách trong thư viện thành văn bản điện tử chỉ trong thời gian ngắn. Nhận diện chữ viết cũng có ứng dụng trong lĩnh vực khoa học hình sự và bảo mật thông tin, như trong việc xác thực chữ ký điện tử.

1.7.2. Nhận dạng giọng nói

Quá trình nhận diện giọng nói cho phép chuyển đổi âm thanh thành văn bản. Điều này giúp người dùng thao tác trên thiết bị một cách nhanh chóng và dễ dàng hơn; ví dụ, thay vì phải gõ một tài liệu, bạn chỉ cần đọc và phần mềm sẽ tự động ghi lại. Đây cũng là bước quan trọng trong việc phát triển giao tiếp giữa con người và robot. Nhận diện giọng nói đặc biệt hữu ích cho những người khiếm thị.

1.7.3. Tổng hợp tiếng nói

Tổng hợp tiếng nói là quá trình chuyển đổi văn bản thành âm thanh. Thay vì tự đọc sách hoặc nội dung từ một trang web, hệ thống sẽ tự động đọc cho bạn. Tương tự như nhận diện giọng nói, tổng hợp tiếng nói hỗ trợ người khiếm thị, nhưng nó cũng là bước cuối cùng trong giao tiếp giữa robot và con người.

1.7.4. Dịch tự động

Như tên gọi, dịch tự động là phần mềm chuyển đổi văn bản từ ngôn ngữ này sang ngôn ngữ khác. Một ví dụ điển hình cho tiếng Việt là phần mềm Evtrans của Softex, cho phép dịch tự động giữa tiếng Anh và tiếng Việt. Phần mềm này từng được trang web vdict.com mua bản quyền và là một trong những ứng dụng đầu tiên trên mạng. Vào tháng 10 năm 2008, hai công ty Lạc Việt và Google đã tham gia vào lĩnh vực dịch tự động cho tiếng Việt, và sau đó Xalo.vn cũng cung cấp dịch vụ tương tự.

1.7.5. Tìm kiếm thông tin

Chương trình tìm kiếm thông tin cho phép người dùng đặt câu hỏi và tự động tìm kiếm nội dung phù hợp nhất. Với sự phát triển của internet, việc tiếp cận thông tin trở nên dễ dàng hơn, nhưng thách thức là tìm ra thông tin chính xác và đáng tin cậy giữa một khối lượng lớn dữ liệu. Các công cụ tìm kiếm như Google hay Yahoo hiện nay chủ yếu dựa vào tần suất từ khóa và các tiêu chí khác để phân tích nội dung, dẫn đến nhiều trường hợp không tìm được câu trả lời chính xác. Đến nay, chưa có công cụ tìm kiếm nào thực sự hiểu ngôn ngữ tự nhiên của con người, ngoại trừ trang www.ask.com, được cho là có khả năng hiểu các câu hỏi đơn giản. Gần đây, trang Wolfram Alpha đã thu hút sự chú ý với khả năng hứa hẹn hiểu ngôn ngữ tự nhiên và cung cấp câu trả lời chính xác.

1.7.6. Tóm tắt văn bản tự động

Tóm tắt văn bản tự động là quá trình rút gọn nội dung của văn bản gốc để tạo ra một bản tóm tắt ngắn gọn, tập trung vào những điểm quan trọng. Có hai phương pháp chính: phương pháp trích xuất, trong đó các câu được lấy nguyên từ văn bản, và phương pháp tóm lược ý, trong đó thông tin chính được truyền đạt bằng ngôn ngữ của người tóm tắt.

1.7.7. Khai phá dữ liệu

Khai phá dữ liệu là quá trình tìm kiếm tri thức mới từ một lượng lớn tài liệu. Việc này tương tự như quá trình học tập và khám phá của con người, và hiện đang trong giai đoạn phát triển. Khi kết hợp với công cụ tìm kiếm, nó cho phép người dùng đặt câu hỏi và tự động tìm ra câu trả lời dựa trên thông tin có sẵn trên web, tương tự như dịch vụ hỏi đáp của Yahoo!.

1.7.8. Truy xuất và trích chọn thông tin

Truy xuất thông tin (Information Retrieval - IR) là quá trình tìm kiếm tài liệu không có cấu trúc (thường là văn bản) để đáp ứng nhu cầu thông tin từ các nguồn lớn. Các hệ thống truy xuất thông tin phổ biến như Google, Yahoo, hay Bing cho phép người dùng nhập câu truy vấn bằng ngôn ngữ tự nhiên và nhận lại danh sách tài liệu phù hợp.

Trích chọn thông tin (Information Extraction) là quá trình nhận diện các thực thể, mối quan hệ và sự kiện trong văn bản. Khác với truy xuất thông tin, trích chọn thông tin cung cấp chính xác thông tin mà người dùng cần, có thể liên quan đến con người, địa điểm, tổ chức, thời gian, hoặc các thông tin cụ thể khác.

1.7.9. Trả lời câu hỏi

Trả lời câu hỏi (QA) là khả năng tự động cung cấp câu trả lời cho các câu hỏi của con người bằng cách truy xuất thông tin từ một tập hợp tài liệu. Một hệ thống QA thường bao gồm ba mô-đun: mô-đun xử lý truy vấn, mô-đun xử lý tài liệu và mô-đun xử lý câu trả lời.

1.7.10. Dịch máy

Dịch máy (Machine Translation - MT) là việc sử dụng máy tính để tự động hóa quá trình dịch giữa các ngôn ngữ. Các phương pháp dịch máy phổ biến bao gồm dịch dựa trên ví dụ, dịch dựa trên luật và dịch thống kê. Nghiên cứu gần đây tập trung vào dịch máy thống kê do nhiều ưu điểm của nó so với các phương pháp khác, với các mô hình như dịch dựa trên từ, cú pháp, cụm từ và cụm từ phân cấp.

CHƯƠNG II: XỬ LÝ NGÔN NGỮ TỰ NHIÊN

2.1. Thách thức cơ bản trong xử lý ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên (NLP) bao gồm hai lĩnh vực chính: hiểu ngôn ngữ tự nhiên (Natural Language Understanding - NLU) và sinh ngôn ngữ tự nhiên (Natural Language Generation - NLG). Trong đó, NLU được chia thành bốn bước quan trọng như sau:

2.1.1. Phân tích hình vị

Trong ngôn ngữ học, hình vị là đơn vị ngữ pháp nhỏ nhất có nghĩa trong một ngôn ngữ. Nghiên cứu về hình vị được gọi là hình vị học. Hình vị có thể độc lập hoặc phụ thuộc, trong khi từ luôn có khả năng đứng độc lập.

Ví dụ, trong tiếng Anh, từ "chair" có nghĩa là "cái ghế" và là một hình vị độc lập. Khi chuyển sang số nhiều với "chairs", ta có hai hình vị: "chair" (hình vị độc lập) và "-s" (hình vị phụ thuộc). Hình vị "-s" không thể đứng một mình, nhưng vẫn mang ý nghĩa.

Mối quan hệ giữa hình vị và từ là yếu tố quan trọng trong việc phân loại ngôn ngữ. Ví dụ, trong tiếng Việt, một hình vị thường là một từ, trong khi trong tiếng Nhật, một từ có thể chứa nhiều hình vị. Phân tích hình vị liên quan đến việc nhận diện, phân tích và mô tả cấu trúc của các hình vị trong một ngôn ngữ cụ thể, bao gồm các đơn vị như từ gốc, biên từ, phụ tố, và từ loại. Hai bài toán điển hình trong lĩnh vực này là tách từ (word segmentation) và gán nhãn từ loại (POS tagging).

2.1.2. Phân tích cú pháp

Phân tích cú pháp là quá trình phân tích một chuỗi biểu tượng, có thể là ngôn ngữ tự nhiên hoặc ngôn ngữ máy tính, theo quy tắc của văn phạm hình thức. Các loại văn phạm thường được sử dụng trong phân tích cú pháp bao gồm Văn phạm phi ngữ cảnh (CFG), Văn phạm danh mục kết nối (CCG), và Văn phạm phụ thuộc (DG). Đầu vào của quá trình này là một câu với chuỗi từ và nhãn từ loại, và đầu ra là một cây phân tích thể hiện cấu trúc cú pháp của câu. Một số thuật toán phổ biến trong phân tích cú pháp bao gồm CKY, Earley, Chart, và GLR.

2.1.3. Phân tích ngữ nghĩa

Quá trình phân tích ngữ nghĩa liên quan đến việc liên kết cấu trúc ngữ nghĩa từ cấp độ cụm từ, mệnh đề, câu cho đến toàn bộ bài viết, nhằm tìm ra ý nghĩa của ngôn từ. Phân tích ngữ nghĩa bao gồm hai cấp độ: ngữ nghĩa từ vựng, liên quan đến ý nghĩa của các từ thành phần, và ngữ nghĩa thành phần, liên quan đến cách các từ kết hợp để tạo ra những nghĩa rộng hơn.

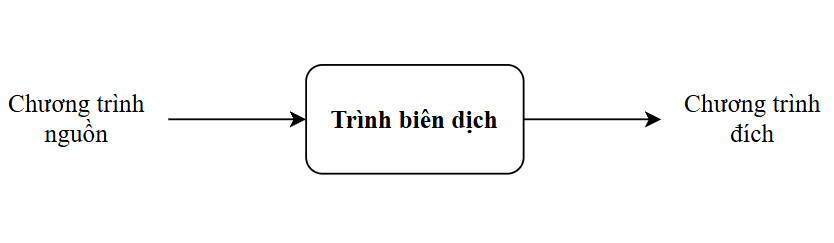
2.1.4. Phân tích diễn ngôn

Ngữ dụng học nghiên cứu mối quan hệ giữa ngôn ngữ và ngữ cảnh sử dụng. Ngữ cảnh này bao gồm danh tính của người hoặc vật, và do đó, ngữ dụng học cũng nghiên cứu cách ngôn ngữ được sử dụng để đề cập hoặc tái đề cập đến người hoặc vật. Ngữ cảnh sử dụng bao gồm cả ngữ cảnh diễn ngôn, vì vậy ngữ dụng học cũng xem xét cách thức cấu tạo diễn ngôn và cách người nghe hiểu người nói.

Một khía cạnh quan trọng khác của NLP là sinh ngôn ngữ tự nhiên (NLG). Đây là nhiệm vụ tạo ra ngôn ngữ tự nhiên từ một hệ thống máy biểu diễn, như cơ sở tri thức hoặc dạng biểu diễn logic. NLG đóng vai trò thiết yếu trong nhiều ứng dụng của NLP, bao gồm sinh hội thoại, tương tác người-máy, dịch thuật máy, và tóm tắt văn bản tự động.

2.2. Qui trình xử lý ngôn ngữ tự nhiên

Để máy tính có thể hiểu và thực thi một chương trình viết bằng ngôn ngữ cấp cao, cần có một trình biên dịch để chuyển đổi chương trình đó sang ngôn ngữ đích. Chương này sẽ cung cấp cái nhìn tổng quan về cấu trúc của trình biên dịch và mối liên hệ của nó với các thành phần khác như bộ tiền xử lý, bộ tải và soạn thảo liên kết. Cấu trúc của trình biên dịch được mô tả theo một mô hình khái niệm, bao gồm các giai đoạn: Phân tích hình thái, Phân tích từ vựng, Phân tích cú pháp, Sinh mã trung gian, Tối ưu mã và Sinh mã đích. Nói một cách đơn giản, trình biên dịch là một chương trình đọc một chương trình viết bằng ngôn ngữ nguồn và dịch nó thành một chương trình tương đương ở ngôn ngữ đích. Một phần quan trọng trong quá trình này là ghi nhận các lỗi trong chương trình nguồn để thông báo cho người lập trình.



Hình 2.1: Một trình biên dịch

2.2.1. Phân tích từ vựng (Lexical Analysis)

Giai đoạn này sẽ đọc chương trình nguồn từ trái sang phải (quét nguyên liệu - scanning) để tách ra thành các thẻ từ (token).

*Ví dụ 1:* Quá trình phân tích từ vựng cho câu lệnh **position := initial + rate \* 60** sẽ tách thành các token như sau:

* 1. Danh biểu **position**
  2. Ký hiệu phép gán **:=**
  3. Danh biểu **initial**
  4. Ký hiệu phép cộng: +
  5. Danh biểu **rate**
  6. Ký hiệu phép nhân: **\***
  7. Số **60**

Trong quá trình này, các khoảng trắng sẽ bị bỏ qua.

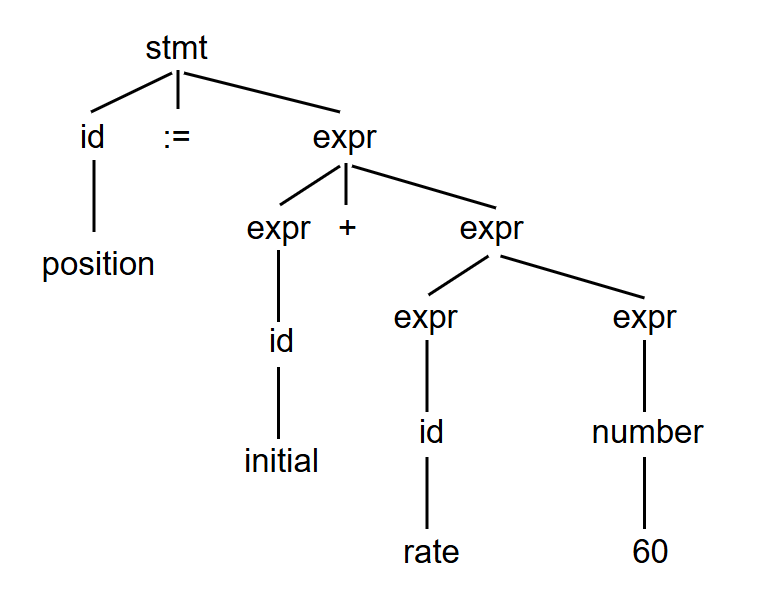
2.2.2. Phân tích cú pháp (Syntax Analysis)

Dãy các từ sẽ được biến đổi thành các cấu trúc thể hiện sự liên kết giữa các từ. Một số dãy từ có thể bị loại bỏ do vi phạm các quy tắc văn phạm. Giai đoạn này nhóm các thẻ từ thành các ngữ đoạn văn phạm (grammatical phrase), thường được biểu diễn bằng cây phân tích cú pháp (parse tree).

*Ví dụ 2:* Giả sử ngôn ngữ được đặc tả bởi các luật sinh sau:

* + **Stmt → id := expr**
  + **expr → expr + expr | expr \* expr | id | number**

Với câu nhập: **position := initial + rate \* 60**, cây phân tích cú pháp sẽ được xây dựng như sau:



Hình 2.2: Một cây phân tích cú pháp

Cấu trúc phân cấp của một chương trình thường được diễn tả bởi quy luật đệ quy.

*Ví dụ 3:*

* 1. Danh biểu (identifier) là một biểu thức (expr).
  2. Số (number) là một biểu thức.
  3. Nếu **expr1** và **expr2** là các biểu thức thì:

**expr1 + expr2**, **expr1 \* expr2**, **(expr)**

* 1. Câu lệnh (statement) cũng có thể định nghĩa đệ quy:

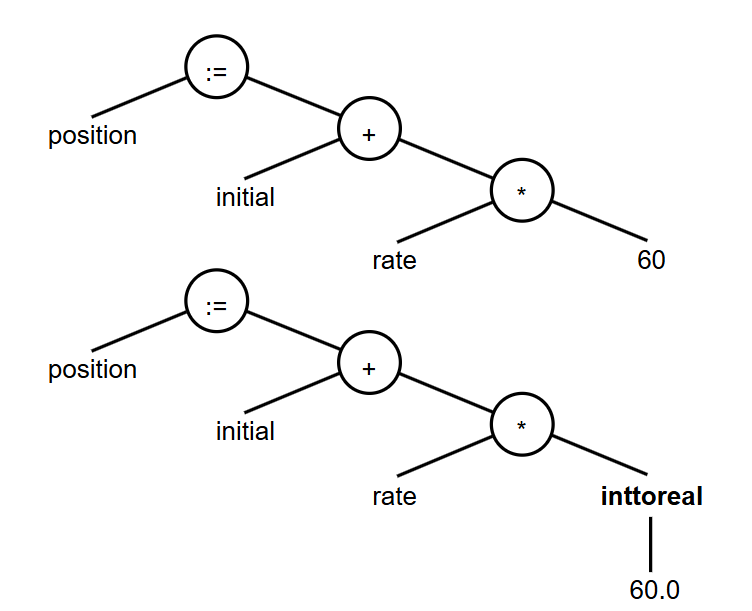
Nếu **id1** là một danh biểu và **expr2** là một biểu thức thì **id1 := expr2** là một lệnh (stmt).

Nếu **expr1** là một biểu thức và **stmt2** là một lệnh thì **while (expr1) do stmt2** và **if (expr1) then stmt2** đều là các lệnh.

2.2.3. Phân tích ngữ nghĩa (Semantic Analysis)

Giai đoạn này thêm ngữ nghĩa vào các cấu trúc được tạo ra bởi bộ phân tích cú pháp. Nó kiểm tra xem chương trình nguồn có chứa lỗi ngữ nghĩa hay không và tập hợp thông tin về kiểu cho giai đoạn sinh mã sau. Một phần quan trọng trong giai đoạn này là kiểm tra kiểu (type checking) và ép chuyển đổi kiểu.

*Ví dụ 4:* Trong biểu thức **position := initial + rate \* 60**, các danh biểu được khai báo là **real**, trong khi **60** là số **integer**, do đó trình biên dịch sẽ chuyển đổi số nguyên **60** thành số thực **60.0**.

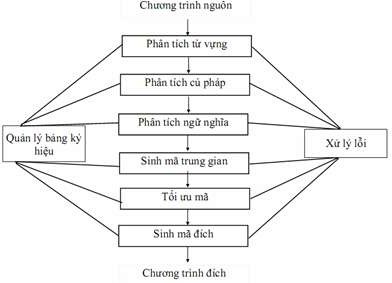


Hình 2.3: Chuyển đổi kiểu trên cây phân tích cú pháp

2.2.4. Các giai đoạn của trình biên dịch

Một trình biên dịch được chia thành các giai đoạn, mỗi giai đoạn chuyển chương trình nguồn từ một dạng biểu diễn này sang một dạng biểu diễn khác.

*Ví dụ:* Một cách phân rã điển hình trình biên dịch được trình bày trong hình 2.4:



Hình 2.4: Các giai đoạn của một trình biên dịch

Quá trình quản lý bảng ký hiệu và xử lý lỗi diễn ra liên tục trong tất cả các giai đoạn của trình biên dịch. Mặc dù các giai đoạn đã được đề cập trước đó thực hiện theo một trình tự logic, trong thực tế, nhiều hoạt động của các giai đoạn có thể được nhóm lại với nhau. Thông thường, chúng được chia thành hai phần chính: phần đầu (Front End) và phần sau (Back End).

***a. Phần Đầu (Front End)***

Phần đầu bao gồm các giai đoạn hoặc các phần giai đoạn mà phụ thuộc nhiều vào ngôn ngữ nguồn và thường không phụ thuộc vào máy đích. Các giai đoạn chính trong phần đầu bao gồm:

* Phân tích từ vựng
* Phân tích cú pháp
* Phân tích ngữ nghĩa
* Sinh mã trung gian

Một phần của công việc tối ưu hóa mã cũng được thực hiện trong giai đoạn này. Phần đầu cũng đảm nhận việc xử lý lỗi phát sinh trong từng giai đoạn, đảm bảo rằng các vấn đề được phát hiện và thông báo kịp thời cho người lập trình.

***b. Phần Sau (Back End)***

Phần sau bao gồm các phần của trình biên dịch mà phụ thuộc vào máy đích và thường không phụ thuộc vào ngôn ngữ nguồn, mà thay vào đó là ngôn ngữ trung gian. Trong phần này, chúng ta sẽ gặp một số vấn đề liên quan đến tối ưu hóa mã, phát sinh mã đích, cùng với việc xử lý lỗi và các thao tác trên bảng ký hiệu.

Phần sau đóng vai trò quan trọng trong việc chuyển đổi mã trung gian thành mã có thể thực thi trên máy đích, đồng thời tối ưu hóa hiệu suất của mã được sinh ra.

2.2.5. Một số phương pháp phân tích cú pháp

2.2.5.1. Top-down

- Phân tích bắt đầu từ gốc của cây cú pháp và di chuyển xuống dưới, từ trái qua phải.

- Khi gặp một từ (terminal), quá trình sẽ tiếp tục phân tích nút tiếp theo.

- Nếu không tìm thấy sự tương ứng với từ đầu vào, thuật toán sẽ quay lại (backtrack) để thử các nhánh khác.

2.2.5.2. Bottom-up

- Đây là một dạng của các hành động shift-reduce.

- Khi gặp vế phải của một quy tắc trong văn phạm, thuật toán sẽ thu gọn (reduce) thành vế trái.

- Nếu không thể tiếp tục phân tích, thuật toán sẽ quay lui (backtrack) để tìm kiếm các khả năng khác.

2.2.5.3. CYK (Cocke-Younger-Kasami)

Văn phạm chuẩn Chomsky (Chomsky Normal Form - NCF)

Các quy tắc trong văn phạm này có thể được phân loại thành hai dạng: A 🡪 B C, A 🡪 a.

*Ví dụ:*

S 🡪 X Y

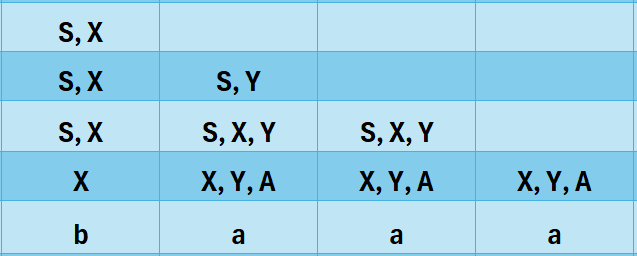
X 🡪 X A | a | b

Y 🡪 A Y | a

A 🡪 a

Phân tích: “babaa” 🡪 không sinh ra câu

“baaa” 🡪 sinh ra câu



Hình 2.5: Phân tích ví dụ CYK

Xác định các đặc điểm sau đây:

a. Sinh ra giá trị một nút như thế nào?

A[i,j] <- ? + ?

b. Lưu lại đường đi như thế nào để sinh lại cây.

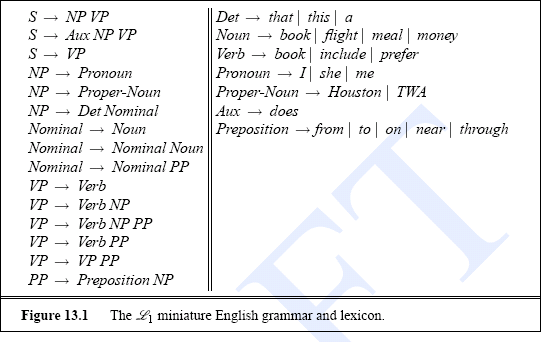
Tính nhập nhằng: Một A[,] có thể có nhiều tag, mỗi tag lại được dẫn xuất bằng nhiều cách.

c. Tại sao thuật toán CYK lại cần văn phạm dạng chuẩn Chomsky.

Phân tích câu:

“book that flight”

“book the flight through Houston”



Hình 2.6: Ngữ pháp tiếng Anh thu nhỏ

Chuyển từ văn phạm CFG sang văn phạm dạng chuẩn Chomsky

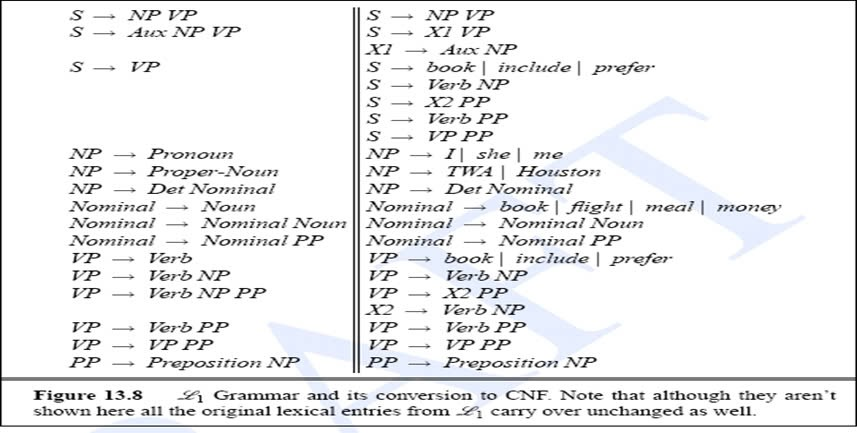
a. A 🡪 BCD

A -> X D

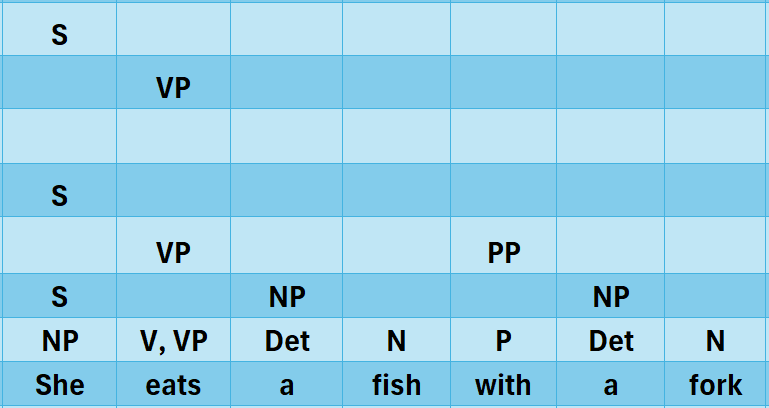
X -> B C

b. Bỏ luật dạng A -> B

Với mọi B 🡪 α, sinh luật A 🡪 α

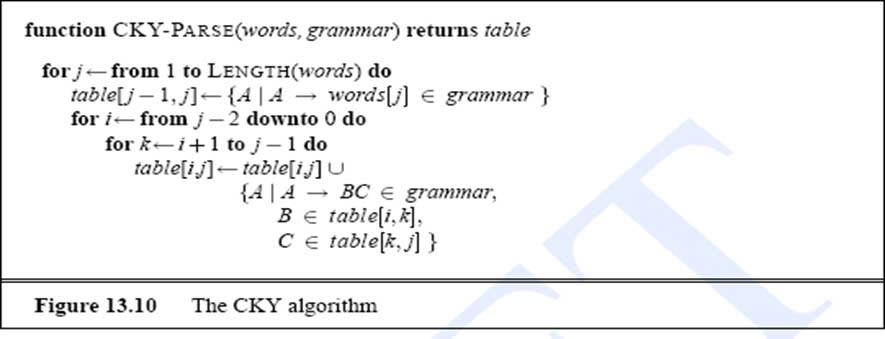


Hình 2.7: Ngữ pháp và cách chuyển đổi sang CNF



Hình 2.8: Thử sinh ra một văn phạm tương ứng

*Thuật toán parsing CYK*



Hình 2.9: Thuật toán parsing CYK

**Đặc điểm:**

- Có thể chuyển mọi văn phạm dạng CFG về dạng chuẩn Chomsky Searching theo kiểu Bottom-up.

- Độ phức tạp phân tích là O(n3).

- Thuật toán là một dạng của *dynamic programming.*

- Có thể mở rộng thuật toán CYK để phân tích văn phạm xác suất.

2.3. Các thì trong tiếng Anh

Bảng 2.1: Các thì trong tiếng Anh

|  |  |
| --- | --- |
| **THÌ** | **CÔNG THỨC** |
| **Hiện tại (Present)** | |
| Hiện tại đơn | S + V1 (số ít, ngôi thứ 3 + es/es) + O |
| Hiện tại tiếp diễn | S + am/is/are + Ving + O |
| Hiện tại hoàn thành | S + has/have + V3 + O |
| Hiện tại hoàn thành tiếp diễn | S + has/have + been + Ving + O |
| **Quá khứ (Past)** | |
| Quá khứ đơn | S + V2 + O |
| Quá khứ tiếp diễn | S + was/were + Ving + O |
| Quá khứ hoàn thành | S + had + V3 |
| Quá khứ hoàn thành tiếp diễn | S + had been + Ving + O |
| **Tương lai (Future)** | |
| Tương lai đơn | S + will/shall + V1 + O |
| Tương lai tiếp diễn | S + will+shall + be + Ving + O |
| Tương lai hoàn thành | S + will/shall + have + V3 |
| Tương lai hoàn thành tiếp diễn | S + will/shall + have been + Ving |

2.4. Cách sử dụng một số thì

2.4.1. Thì hiện tại đơn (The Simple Present Tense)

2.4.1.1. Hình thức

Bảng 2.2: Cách sử dụng thì Hiện tại đơn

|  |  |
| --- | --- |
| **THỂ** | **CÔNG THỨC** |
| Khẳng định | **Verb: S + V(s/es) + O**  *Nếu chủ ngữ là ngôi thứ 3 số ít (He, She, It, hoặc một danh từ), động từ phải thêm “S” hoặc “ES”.*  **Tobe: S + am/is/are + N/Adj** |
| Phủ định | **Verb: S + do/does + not + V + O**  *“Does not” được sử dụng khi chủ ngữ là ngôi thứ 3 số ít, khi đó động từ ở dạng nguyên thể (không thêm “S” hoặc “ES”).*  **Tobe: S + am/is/are + not + N/Adj** |
| Nghi vấn | **Verb: Do/Does + S + V + O?**  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + do/does * Phủ định: No, S + don’t/doesn’t   **Tobe: Am/Is/Are + S + N/Adj?** |
| Câu hỏi Wh- | **Verb: Wh- + do/does + S + V + O?**  **Tobe: (+)Wh- + am/are/is + S + O?**  **(-)Wh- + am/are/is + not + S + O?** |

2.4.1.2. Cách sử dụng

* Diễn tả một sự thật hiển nhiên.
* Một hành động xảy ra hàng ngày, có tính lặp đi lặp lại.
* Diễn tả một hành động ở tương lai (thường dùng với các động từ chỉ sự chuyển động như: arrive, leave, return…).

*Ví dụ: She leaves tomorrow.*

***Dấu hiệu nhận biết:*** Always, usually, never, every (ngày, tuần, tháng, năm), ….

2.4.2. Thì hiện tại tiếp diễn (The present continuous/progressive tense)

2.4.2.1. Hình thức

Bảng 2.3: Cách sử dụng thì Hiện tại tiếp diễn

|  |  |
| --- | --- |
| **THỂ** | **CÔNG THỨC** |
| Khẳng định | **S + am/is/are + V\_ing + O** |
| Phủ định | **S + am/is/are + not + V\_ing + O**  *Am not = ’m not, is not = isn’t, are not = aren’t.* |
| Nghi vấn | **Am/Is/Are + S + V\_ing + O?**  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + am/are/is * Phủ định: No, S + ‘m not/isn’t/aren’t |
| Câu hỏi Wh- | **Wh- + am/is/are + S + V\_ing + O?** |

2.4.2.2. Cách sử dụng

* Diễn tả một hành động đang xảy ra tại thời điểm nói.

*Ví dụ: We are learning English now.*

* Một hành động xảy ra ở tương lai gần.

*Ví dụ: He is watching television tonight.*

* Một hành động được lặp đi lặp lại nhiều lần, gây bực mình (thường có trạng từ “always”).

*Ví dụ: She leaves tomorrow.*

***Dấu hiệu nhận biết:*** Now, at the moment, right now, currently, today, this week, this month, ….

2.4.3. Thì hiện tại hoàn thành (The Present Perfect Tense)

2.4.3.1. Hình thức

Bảng 2.4: Cách sử dụng thì Hiện tại hoàn thành

|  |  |
| --- | --- |
| **THỂ** | **CÔNG THỨC** |
| Khẳng định | **Verb: S + have/has + V3/ed + O**  *Have = ’ve, has = ’s Nếu chủ ngữ là ngôi thứ 3 số ít thì dùng “has”.*  **Tobe: S + have/has been + N/Adj** |
| Phủ định | **Verb: S + have/has + not + V3/ed + O**  **Tobe: S + have/has + not + been+ N/Adj** |
| Nghi vấn | **Verb: Have/Has + S + V3/ed + O?**  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + have/has * Phủ định: No, S + haven’t/hasn’t   **Tobe: Have/Has + S + been + N/Adj?** |
| Câu hỏi Wh- | **Wh- + am/is/are + S + V\_ing + O?** |

2.4.3.2. Cách sử dụng

* Diễn tả một hành động vừa mới xảy ra (thường có trạng từ “just”).  
  *Ví dụ: I have just bought this car.*
* Diễn tả một hành động xảy ra trong quá khứ không xác định thời gian (thường có trạng từ “already”).

*Ví dụ: He has already read that book.*

* Diễn tả một hành động bắt đầu ở quá khứ và vẫn còn tiếp tục ở hiện tại.  
  Các trạng từ chỉ thời gian thường được dùng: ever, never, so far, since (điểm thời gian), for (khoảng thời gian)…

*Ví dụ: I have never driven a car. They have lived here since 1998.*

***Dấu hiệu nhận biết:*** Just, already, yet, ever, never, for, since, so far, up to now.

2.4.4. Thì hiện tại hoàn thành tiếp diễn (The Present Perfect continuousTense)

2.4.4.1. Hình thức

Bảng 2.5: Cách sử dụng thì Hiện tại hoàn thành tiếp diễn

|  |  |
| --- | --- |
| **THỂ** | **CÔNG THỨC** |
| Khẳng định | **S + have/has + been + V\_ing + O** |
| Phủ định | **S + have/has + not + been + V\_ing + O** |
| Nghi vấn | **Have/Has + S + been + V\_ing + O?**  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + have/has * Phủ định: No, S + haven’t/hasn’t |
| Câu hỏi Wh- | **Wh- + have/has + S + been + V\_ing + O?** |

2.4.4.2. Cách sử dụng

* Diễn tả một hành động bắt đầu và còn liên tục đến hiện tại, chấm dứt ở hiện tại hoặc có thể kéo dài đến tương lai.

*Ví dụ: I have been waiting for you for a long time.*

* Lý do xảy ra ngay khi nói.

*Ví dụ: Your eyes are very red. Have you been crying?*

***Dấu hiệu nhận biết:*** For, since, all day, all week, recently, lately.

2.4.5. Thì quá khứ đơn (The Simple Past Tense)

2.4.5.1. Hình thức

Bảng 2.6: Cách sử dụng thì Quá khứ đơn

|  |  |
| --- | --- |
| **THỂ** | **CÔNG THỨC** |
| Khẳng định | **Verb: S + V2/ed + O**  *Nếu là động từ có quy tắc thì thêm “ED” vào sau động từ thường, nếu là động từ bất quy tắc thì sử dụng động từ ở cột 2 trong bảng động từ bất quy tắc.*  **Tobe: S + was/were + N/Adj** |
| Phủ định | **Verb: S + did+ not + V + O**  *Khi có trợ động từ “didn’t”, động từ theo sau trở về nguyên thể.*  **Tobe: S + was/were + not + N/Adj** |
| Nghi vấn | **Verb: Did + S + V + O?**  *Khi có trợ động từ “Did”, động từ ở dạng nguyên thể.*  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + did * Phủ định: No, S + didn’t   **Tobe: Was/Were + S + N/Adj?** |
| Câu hỏi Wh- | **Wh- + did + S + V + O?** |

2.4.5.2. Cách sử dụng

* Diễn tả một hành động xảy ra tại một thời điểm xác định trong quá khứ và đã chấm dứt.
* Diễn tả thói quen trong quá khứ.

*Ví dụ: She often played badminton when she was young.*

* Diễn tả các hành động xảy ra kế tiếp nhau trong quá khứ.

*Ví dụ: She came in, sat down and said nothing.*

***Dấu hiệu nhận biết:*** Yesterday, last (night, week, month, year), ago, in (year), when.

2.4.6. Thì quá khứ tiếp diễn (The Past continuous Tense)

2.4.6.1. Hình thức

Bảng 2.7: Cách sử dụng thì Quá khứ tiếp diễn

|  |  |
| --- | --- |
| **THỂ** | **CÔNG THỨC** |
| Khẳng định | **S + was/were + V\_ing + O**  *Was: dùng cho ngôi I và ngôi thứ 3 số ít.* |
| Phủ định | **S + was not/were not + V\_ing + O**  *Was not = wasn’t, were not = weren’t.* |
| Nghi vấn | **Was/Were + S + V\_ing + O?**  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + was/were * Phủ định: No, S + wasn’t/weren’t |
| Câu hỏi Wh- | **Wh- + was/were + S + V\_ing + O?** |

2.4.6.2. Cách sử dụng

* Diễn tả một hành động đang diễn ra tại một thời điểm trong quá khứ.

*Ví dụ: I was reading a book at 8 o’clock last night.*

* Diễn tả một hành động đang xảy ra ở quá khứ thì bị một hành động khác cắt ngang (hành động cắt ngang dùng thì quá khứ đơn).

*Ví dụ: We were watching TV when the light went out.*

* Một sự việc xảy ra và liên tục trong quá khứ.

*Ví dụ: I was sleeping all day yesterday.*

* Chỉ hai hành động xảy ra song song nhau trong quá khứ.

*Ví dụ: My father was reading the newspaper while my mother was listening to music.*

**Dấu hiệu nhận biết:** While, when, at that time, at (a specific time in the past).

2.4.7. Thì quá khứ hoàn thành (The Past Perfect Tense)

2.4.7.1. Hình thức

Bảng 2.8: Cách sử dụng thì Quá khứ hoàn thành

|  |  |
| --- | --- |
| **THỂ** | **CÔNG THỨC** |
| Khẳng định | **Verb: S + had V3 + O**  **Tobe: S + had + N/Adj** |
| Phủ định | **Verb: S + had + not + V3 + O**  **Tobe: S + had + not + been + N/Adj** |
| Nghi vấn | **Verb: Had + S + V3?**  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + had * Phủ định: No, S + hadn’t   **Tobe: Had + S + been + N/Adj?** |
| Câu hỏi Wh- | **Wh- + had + S + V3/ed + O?** |

2.4.7.2. Cách sử dụng

* Diễn tả một hành động đã xảy ra trước một thời điểm hoặc một hành động khác trong quá khứ.

*Ví dụ: She had finished her homework before she went out.*

* Diễn tả một hành động đã xảy ra trước một hành động khác trong quá khứ.

*Ví dụ: They had left when I arrived.*

***Dấu hiệu nhận biết:*** Before, after, by the time, when (trong ngữ cảnh quá khứ).

2.4.8. Thì quá khứ hoàn thành tiếp diễn (The Past Perfect Continuous Tense)

2.4.8.1. Hình thức

Bảng 2.9: Cách sử dụng thì Quá khứ tiếp diễn

|  |  |
| --- | --- |
| **THỂ** | **CÔNG THỨC** |
| Khẳng định | **S + had + been + V\_ing + O** |
| Phủ định | **S + had + not + been + V\_ing + O** |
| Nghi vấn | **Verb: Had + S + been + V\_ing + O?**  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + had * Phủ định: No, S + hadn’t |
| Câu hỏi Wh- | **Wh- + had + S + been + V\_ing + O?** |

2.4.8.2. Cách sử dụng

* Diễn tả một hành động đã xảy ra và kéo dài liên tục đến một thời điểm trong quá khứ.

*Ví dụ: They had been waiting for two hours when the bus finally arrived.*

* Diễn tả một hành động đã xảy ra liên tục trước một thời điểm hoặc một hành động khác trong quá khứ.

*Ví dụ: She was tired because she had been studying all night.*

***Dấu hiệu nhận biết:*** For, since, all day, all week, before (trong ngữ cảnh quá khứ).

2.4.9. Thì tương lai đơn (The Simple Future Tense)

2.4.9.1. Hình thức

Bảng 2.10: Cách sử dụng thì Tương lai đơn

|  |  |
| --- | --- |
| **THỂ** | **CÔNG THỨC** |
| Khẳng định | **Verb: S + will/shall + V + O**  *Shall được dùng cho ngôi I và We. Trong văn nói và trong tiếng Anh ngày nay, người ta sử dụng “will” cho tất cả các ngôi. ‘ll: viết tắt của Shall và Will.*  **Tobe: S + will be + N/Adj** |
| Phủ định | **Verb: S + will not + shall not + V + O**  *will not = won’t, shall not = shan’t.*  **Tobe: S + will + not + be+ N/Adj** |
| Nghi vấn | **Verb: Will/Shall + S + V+ O?**  *Khi có trợ động từ “Did”, động từ ở dạng nguyên thể.*  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + will/shall * Phủ định: No, S + won’t/shan’t   **Tobe: Will + S + be + N/Adj?**  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + will be + N/Adj * Phủ định: No, S + won’t be + N/Adj |
| Câu hỏi Wh- | **Wh- + will + S + V + O?** |

2.4.9.2. Cách sử dụng

* Diễn tả một hành động sẽ xảy ra tại một thời điểm nào đó trong tương lai.

*Ví dụ: She’ll be 20 on next Thursday.*

* Diễn tả thói quen trong tương lai.

*Ví dụ: He will go for a walk after dinner.*

* Diễn tả một việc sẽ quyết định làm ngay lúc nói.

*Ví dụ: What would you like to drink? I’ll have a mineral water.*

***Dấu hiệu nhận biết:*** Tomorrow, next (week, month, year), in (a specific time in the future), soon, later, in the future.

2.4.10. Thì tương lai tiếp diễn (The Continuous Future Tense)

2.4.10.1. Hình thức

Bảng 2.11: Cách sử dụng thì Tương lai tiếp diễn

|  |  |
| --- | --- |
| **THỂ** | **CÔNG THỨC** |
| Khẳng định | **S + will + be + V\_ing + O** |
| Phủ định | **S + will+ not + be + V\_ing + O** |
| Nghi vấn | **Will + S + be + V\_ing + O?**  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + will * Phủ định: No, S + won’t |
| Câu hỏi Wh- | **Wh- + will + S + V + O?** |

2.4.10.2. Cách sử dụng

* Diễn tả một hành động sẽ đang diễn ra tại một thời điểm xác định trong tương lai.

*Ví dụ: I will be studying at 8 PM tomorrow.*

* Diễn tả một hành động sẽ xảy ra trong tương lai và kéo dài trong một khoảng thời gian.  
  *Ví dụ: They will be traveling for two weeks next month.*

***Dấu hiệu nhận biết:*** At this time tomorrow, at (a specific time in the future), next (week, month, year).

2.4.11. Thì tương lai hoàn thành (The Perfect Future Tense)

2.4.11.1. Hình thức

Bảng 2.12: Cách sử dụng thì Tương lai hoàn thành

|  |  |
| --- | --- |
| **THỂ** | **CÔNG THỨC** |
| Khẳng định | **Verb: S + will + have + V3/ed + O**  **Tobe: S + will have been + N/Adj** |
| Phủ định | **Verb: S + will+ not + have + V3/ed + O**  **Tobe: S + will+ not + have been + N/Adj** |
| Nghi vấn | **Verb: Will + S + have + V3/ed + O?**  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + will * Phủ định: No, S + won’t   **Tobe: Will+ S + have been + N/Adj?**  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + will have been + N/Adj * Phủ định: No, S + won’t have been + N/Adj |
| Câu hỏi Wh- | **Wh- + will+ S + have + V3/ed + O?** |

2.4.11.2. Cách sử dụng

* Diễn tả một hành động sẽ hoàn thành trước một thời điểm hoặc một hành động khác trong tương lai.

*Ví dụ: By next year, I will have graduated from university.*

* Diễn tả một hành động sẽ xảy ra và hoàn thành trước một thời điểm trong tương lai.

*Ví dụ: They will have finished the project by Friday.*

***Dấu hiệu nhận biết:*** By, by the time, before (trong ngữ cảnh tương lai).

2.4.12. Thì tương lai hoàn thành tiếp diễn (The Perfect Continuous Future Tense)

2.4.12.1. Hình thức

Bảng 2.12: Cách sử dụng thì Tương lai hoàn thành tiếp diễn

|  |  |
| --- | --- |
| **THỂ** | **CÔNG THỨC** |
| Khẳng định | **S + will + have + been + V\_ing + O** |
| Phủ định | **S + will+ not + have + been + V\_ing + O** |
| Nghi vấn | **Will + S + have + been + V\_ing + O?**  *Câu trả lời ngắn:*   * Khẳng định: Yes, S + will * Phủ định: No, S + won’t |
| Câu hỏi Wh- | **Wh- + will + S + have + been + V\_ing + O?** |

2.4.12.2. Cách sử dụng

* Diễn tả một hành động sẽ xảy ra và kéo dài liên tục đến một thời điểm trong tương lai.

*Ví dụ: By next month, I will have been working here for five years.*

* Diễn tả một hành động sẽ xảy ra liên tục trước một thời điểm trong tương lai.

*Ví dụ: They will have been studying for three hours by the time the exam starts.*

***Dấu hiệu nhận biết:*** For, since, by the time (trong ngữ cảnh tương lai).

2.5. Các cấu trúc trong câu

2.5.1. Cấu trúc chung của 1 câu trong Tiếng Anh

Trong tiếng Anh, một câu thường bao gồm các thành phần sau đây:

Ví dụ:

Bảng 2.13: Cấu trúc thường gặp của một câu Tiếng Anh

|  |  |  |  |
| --- | --- | --- | --- |
| **SUBJECT** | **VERB** | **COMPLEMENT** | **MODIFIER** |
| The dog | barked | loudly | at the mailman. |
| She | reads | a novel | every evening. |
| They | traveled | to Japan | last summer. |
| The students | finished | their homework. |  |

***Subject (chủ ngữ):***

* Chủ ngữ là phần của câu thể hiện ai hoặc cái gì thực hiện hành động. Nó thường đứng trước động từ và có thể là một danh từ đơn hoặc một cụm danh từ. Chủ ngữ thường xuất hiện ở đầu câu và ảnh hưởng đến cách chia động từ.
* Lưu ý rằng mọi câu trong tiếng Anh đều cần có chủ ngữ. Trong các câu mệnh lệnh, chủ ngữ thường được ngầm hiểu là người nghe. Ví dụ: “Don't move!” có thể hiểu là “You don’t move!”.
* Ví dụ về chủ ngữ:
  + **Milk is delicious.** (danh từ)
  + **That new, red car is mine.** (cụm danh từ)
* Đôi khi, câu có thể không có chủ ngữ cụ thể, trong trường hợp này, "It" hoặc "There" được sử dụng như chủ ngữ giả.
  + **It is a nice day today.**
  + **There is a fire in that building.**
  + **There were many students in the room.**
  + **It is the fact that the earth goes around the sun.**

***Verb (động từ):***

* Động từ là từ thể hiện hành động hoặc trạng thái của chủ ngữ. Mỗi câu đều cần có ít nhất một động từ, có thể là một từ đơn hoặc một cụm động từ. Cụm động từ bao gồm một hoặc nhiều trợ động từ và một động từ chính.
* Ví dụ về động từ:
  + **I love you.** (hành động)
  + **Chili is hot.** (trạng thái)
  + **I have seen the movie three times before.** (trợ động từ: have; động từ chính: seen)
  + **I am going to Sai Gon tomorrow.** (trợ động từ: am; động từ chính: going)

***Complement (vị ngữ):***

* Vị ngữ là phần của câu chỉ ra đối tượng mà chủ ngữ tác động đến. Tương tự như chủ ngữ, vị ngữ thường là danh từ hoặc cụm danh từ và thường đứng sau động từ. Không phải câu nào cũng có vị ngữ. Vị ngữ trả lời cho câu hỏi "What?" hoặc "Whom?".
* Ví dụ về vị ngữ:
  + **John bought a car yesterday.** (What did John buy?)
  + **Jill wants to drink some water.** (What does she want to drink?)
  + **She saw John at the movie last night.** (Whom did she see at the movie?)

***Modifier (trạng từ):***

* Trạng từ là từ hoặc cụm từ mô tả thời gian, địa điểm hoặc cách thức của hành động. Không phải câu nào cũng có trạng từ. Chúng có thể là cụm giới từ, phó từ hoặc cụm phó từ. Trạng từ trả lời cho các câu hỏi "When?", "Where?" hoặc "How?". Cụm giới từ bắt đầu bằng một giới từ và kết thúc bằng một danh từ (VD: in the morning, on the table,...). Nếu có nhiều trạng từ trong câu, trạng từ chỉ thời gian thường đứng ở cuối.
* Ví dụ về trạng từ:
  + **John bought a book at the bookstore.** (Where did John buy a book?)
  + **She saw John at the movie last night.** (Where did she see John? When did she see him?)
  + **She drives very fast.** (How does she drive?)
* Lưu ý rằng trạng từ thường đứng sau vị ngữ nhưng không phải lúc nào cũng vậy. Tuy nhiên, cụm giới từ không được đặt giữa động từ và vị ngữ.
  + **She drove on the street her new car.** (Sai)
  + **She drove her new car on the street.** (Đúng)

2.5.2. Tổng hợp 100 cấu trúc câu Tiếng Anh

1. **S + V + so + adj/adv + that + S + V**: (quá... đến nỗi mà)
   * The movie was so boring that I fell asleep. (Bộ phim chán đến nỗi tôi đã ngủ gật.)
2. **It + V + such + (a/an) + N(s) + that + S + V**: (quá... đến nỗi mà)
   * It was such a beautiful day that we decided to go for a picnic. (Đó là một ngày đẹp đến nỗi chúng tôi quyết định đi dã ngoại.)
3. **S + V + adj/adv + enough + (for someone) + to do something**: (đủ... cho ai đó làm gì)
   * She is smart enough to solve this problem. (Cô ấy đủ thông minh để giải quyết vấn đề này.)
4. **To prevent/stop + someone/something + from + V-ing**: (ngăn cản ai/cái gì... làm gì..)
   * The teacher stopped the students from talking during the exam. (Giáo viên đã ngăn học sinh nói chuyện trong giờ thi.)
5. **S + find + it + adj to do something**: (thấy ... để làm gì...)
   * I find it easy to make friends. (Tôi thấy việc kết bạn rất dễ dàng.)
6. **To prefer + Noun/V-ing + to + N/V-ing**: (thích cái gì/ làm gì hơn cái gì/ làm gì)
   * I prefer coffee to tea. (Tôi thích cà phê hơn trà.)
7. **Would rather ('d rather) + V (infinitive) + than + V (infinitive)**: (thích làm gì hơn làm gì)
   * I would rather stay home than go out tonight. (Tôi thích ở nhà hơn là ra ngoài tối nay.)
8. **To be/get used to + V-ing**: (quen làm gì)
   * I am used to waking up early. (Tôi đã quen với việc dậy sớm.)
9. **Used to + V (infinitive)**: (thường làm gì trong quá khứ và bây giờ không làm nữa)
   * I used to play soccer every weekend. (Tôi từng chơi bóng đá mỗi cuối tuần.)
10. **To be amazed at = to be surprised at + N/V-ing**: (ngạc nhiên về...)
    * I was surprised at how quickly he finished the task. (Tôi rất ngạc nhiên về việc anh ấy hoàn thành nhiệm vụ nhanh chóng như vậy.)
11. **To be angry at + N/V-ing**: (tức giận về)
    * She was angry at him for being late. (Cô ấy đã tức giận với anh ấy vì đến muộn.)
12. **To be good at/bad at + N/V-ing**: (giỏi về.../kém về...)
    * He is bad at math. (Anh ấy rất kém về toán.)
13. **To waste + time/money + V-ing**: (tốn tiền hoặc thời gian làm gì)
    * I waste a lot of time playing video games. (Tôi tốn rất nhiều thời gian để chơi game.)
14. **To spend + amount of time/money + on + something**: (dành thời gian/tiền bạc vào việc gì...)
    * She spends a lot of money on clothes. (Cô ấy dành nhiều tiền vào quần áo.)
15. **To give up + V-ing/N**: (từ bỏ làm gì/cái gì...)
    * He decided to give up smoking for his health. (Anh ấy quyết định từ bỏ thuốc lá vì sức khỏe của mình.)
16. **To look forward to V-ing**: (mong chờ, mong đợi làm gì)
    * I look forward to meeting you next week. (Tôi mong chờ được gặp bạn vào tuần tới.)
17. **To provide smb with smth**: (cung cấp cho ai cái gì)

* The school provides students with all the necessary materials. (Trường học cung cấp cho học sinh tất cả các tài liệu cần thiết.)

1. **To fail to do smt**: (không làm được cái gì)

* He failed to complete the project on time. (Anh ấy không hoàn thành dự án đúng hạn.)

1. **To succeed in V-ing**: (thành công trong việc làm gì)

* She succeeded in passing the difficult exam. (Cô ấy đã thành công trong việc vượt qua kỳ thi khó.)

1. **To borrow smt from smb**: (mượn cái gì của ai)

* Can I borrow your pen for a moment? (Tôi có thể mượn bút của bạn một chút không?)

1. **To lend smb smt**: (cho ai mượn cái gì)

* I can lend you my book if you want. (Tôi có thể cho bạn mượn cuốn sách của tôi nếu bạn muốn.)

1. **To make smb do smt**: (bắt ai làm gì)

* The coach made the players practice harder. (Huấn luyện viên bắt các cầu thủ tập luyện chăm chỉ hơn.)

1. **CN + be + so + tính từ + that + S + động từ**: (đến mức mà)

* The weather was so nice that we decided to go for a walk. (Thời tiết đẹp đến nỗi chúng tôi quyết định đi dạo.)

1. **CN + be + such + (tính từ) + danh từ + that + CN + động từ**: (quá... đến nỗi mà)

* It was such a fun party that everyone enjoyed it. (Đó là một bữa tiệc vui đến nỗi mọi người đều thích thú.)

1. **It is (very) kind of smb to do smt**: (ai thật tốt bụng khi làm gì)

* It is very kind of you to help me with my homework. (Bạn thật tốt bụng khi giúp tôi làm bài tập về nhà.)

1. **To find it + tính từ + to do smt**: (thấy ... để làm gì)

* I find it challenging to learn new languages. (Tôi thấy việc học ngôn ngữ mới là một thách thức.)

1. **To make sure of smt**: (bảo đảm điều gì)

* I need to make sure of the meeting time. (Tôi cần bảo đảm thời gian họp.)

1. **It takes (smb) + thời gian + to do smt**: (mất (của ai) bao nhiêu thời gian để làm gì)

* It takes me an hour to get to work. (Tôi mất một tiếng để đến nơi làm việc.)

1. **To spend + time/money + on smt**: (dành thời gian/tiền bạc vào cái gì)

* We spend a lot of time on our studies. (Chúng tôi dành nhiều thời gian cho việc học.)

1. **To have no idea of smt**: (không biết về cái gì)

* I have no idea of what you're talking about. (Tôi không biết bạn đang nói về điều gì.)

1. **To advise smb to do smt**: (khuyên ai làm gì)

* I advise you to take the bus instead of driving. (Tôi khuyên bạn nên đi xe buýt thay vì lái xe.)

1. **To plan to do smt**: (dự định làm gì)

* We plan to visit our grandparents this weekend. (Chúng tôi dự định thăm ông bà vào cuối tuần này.)

1. **To invite smb to do smt**: (mời ai làm gì)

* They invited us to their wedding. (Họ đã mời chúng tôi đến đám cưới của họ.)

1. **To offer smb smt**: (mời/đề nghị ai cái gì)

* He offered me a ride to the station. (Anh ấy đã mời tôi đi nhờ đến ga.)

1. **To rely on smb**: (tin cậy, dựa dẫm vào ai)

* You can rely on me for support. (Bạn có thể tin tưởng vào tôi để được hỗ trợ.)

1. **To keep a promise**: (giữ lời hứa)

* She always keeps her promises. (Cô ấy luôn giữ lời hứa của mình.)

1. **To be able to do smt**: (có khả năng làm gì)

* I am able to finish the project by tomorrow. (Tôi có khả năng hoàn thành dự án trước ngày mai.)

1. **To be good at (V-ing)**: (giỏi (làm) cái gì)

* He is good at playing the guitar. (Anh ấy chơi guitar rất giỏi.)

1. **To prefer smt to smt**: (thích cái gì hơn cái gì)

* I prefer chocolate to vanilla. (Tôi thích sô cô la hơn vani.)

1. **To apologize for doing smt**: (xin lỗi ai vì đã làm gì)

* I want to apologize for being late. (Tôi muốn xin lỗi vì đã đến muộn.)

1. **Had (’d) better do smt**: (nên làm gì)

* You had better study for the exam. (Bạn nên học cho kỳ thi.)

1. **Would (’d) rather do smt**: (thà làm gì)

* I would rather go for a walk than stay indoors. (Tôi thà đi dạo còn hơn ở trong nhà.)

1. **Would (’d) rather smb did smt**: (muốn ai làm gì)

* I would rather you stayed home tonight. (Tôi muốn bạn ở nhà tối nay.)

1. **To suggest smb (should) do smt**: (gợi ý ai làm gì)

* I suggest you (should) take a break. (Tôi gợi ý bạn nên nghỉ ngơi.)

1. **To suggest doing smt**: (gợi ý làm gì)

* I suggest going to the beach this weekend. (Tôi gợi ý đi biển vào cuối tuần này.)

1. **Try to do**: (cố làm gì)

* I will try to finish my homework before dinner. (Tôi sẽ cố gắng hoàn thành bài tập trước bữa tối.)

1. **Try doing smt**: (thử làm gì)

* You should try cooking this recipe. (Bạn nên thử nấu món này.)

1. **To need to do smt**: (cần làm gì)

* You need to clean your room. (Bạn cần dọn dẹp phòng của mình.)

1. **To need doing**: (cần được làm)

* The car needs washing. (Chiếc xe cần được rửa.)

1. **To remember doing**: (nhớ đã làm gì)

* I remember meeting her at the party. (Tôi nhớ đã gặp cô ấy tại bữa tiệc.)

1. **To remember to do**: (nhớ làm gì)

* Remember to call your mother. (Hãy nhớ gọi cho mẹ bạn.)

1. **To have smt + PII**: (có cái gì được làm)

* I am going to have my house painted. (Tôi sẽ có nhà mình được sơn lại.)

1. **To have smb do smt**: (thuê ai làm gì)

* I will have the mechanic fix my car. (Tôi sẽ nhờ thợ sửa xe sửa chiếc xe của mình.)

1. **To be busy doing smt**: (bận rộn làm gì)

* I am busy studying for my exams. (Tôi đang bận rộn học cho kỳ thi của mình.)

1. **To mind doing smt**: (phiền làm gì)

* Would you mind closing the window? (Bạn có phiền đóng cửa sổ không?)

1. **To be used to doing smt**: (quen với việc làm gì)

* I am used to waking up early every day. (Tôi đã quen với việc dậy sớm mỗi ngày.)

1. **To stop to do smt**: (dừng lại để làm gì)

* We stopped to take a picture. (Chúng tôi dừng lại để chụp một bức ảnh.)

1. **To stop doing smt**: (thôi không làm gì nữa)

* I stopped eating junk food. (Tôi đã thôi không ăn đồ ăn nhanh nữa.)

1. **Let smb do smt**: (để ai làm gì)

* Let me help you with that. (Để tôi giúp bạn với việc đó.)

1. **To be afraid of + N/V-ing**: (sợ cái gì/làm gì)

* She is afraid of flying. (Cô ấy sợ bay.)

1. **To be responsible for + N/V-ing**: (chịu trách nhiệm về cái gì/làm gì)

* He is responsible for managing the team. (Anh ấy chịu trách nhiệm quản lý đội ngũ.)

1. **To be interested in + N/V-ing**: (quan tâm đến cái gì/làm gì)

* I am interested in learning new languages. (Tôi quan tâm đến việc học ngôn ngữ mới.)

1. **To be tired of + N/V-ing**: (mệt mỏi về cái gì/làm gì)

* I am tired of doing the same thing every day. (Tôi mệt mỏi vì làm những việc giống nhau mỗi ngày.)

1. **To be excited about + N/V-ing**: (hào hứng về cái gì/làm gì)

* I am excited about going on vacation. (Tôi rất hào hứng về việc đi nghỉ.)

1. **To be good for + N**: (tốt cho cái gì)

* Exercise is good for your health. (Tập thể dục tốt cho sức khỏe của bạn.)

1. **To be bad for + N**: (xấu cho cái gì)

* Smoking is bad for your health. (Hút thuốc có hại cho sức khỏe của bạn.)

1. **To be capable of + V-ing**: (có khả năng làm gì)

* She is capable of solving complex problems. (Cô ấy có khả năng giải quyết các vấn đề phức tạp.)

1. **To be familiar with + N/V-ing**: (quen thuộc với cái gì/làm gì)

* I am familiar with this software. (Tôi quen thuộc với phần mềm này.)

1. **To be in charge of + N/V-ing**: (chịu trách nhiệm về cái gì/làm gì)

* She is in charge of the marketing department. (Cô ấy chịu trách nhiệm về bộ phận tiếp thị.)

1. **To be involved in + N/V-ing**: (tham gia vào cái gì/làm gì)

* He is involved in several community projects. (Anh ấy tham gia vào một số dự án cộng đồng.)

1. **To be known for + N/V-ing**: (được biết đến vì cái gì/làm gì)

* This restaurant is known for its delicious food. (Nhà hàng này được biết đến vì món ăn ngon.)

1. **To be used for + N/V-ing**: (được sử dụng cho cái gì/làm gì)

* This tool is used for cutting wood. (Công cụ này được sử dụng để cắt gỗ.)

1. **To be suitable for + N/V-ing**: (phù hợp với cái gì/làm gì)

* This dress is suitable for formal occasions. (Chiếc váy này phù hợp cho các dịp trang trọng.)

1. **To be dedicated to + N/V-ing**: (cống hiến cho cái gì/làm gì)

* She is dedicated to helping others. (Cô ấy cống hiến cho việc giúp đỡ người khác.)

1. **To be committed to + N/V-ing**: (cam kết với cái gì/làm gì)

* He is committed to improving his skills. (Anh ấy cam kết cải thiện kỹ năng của mình.)

1. **To be likely to do smt**: (có khả năng làm gì)

* It is likely to rain tomorrow. (Có khả năng ngày mai sẽ mưa.)

1. **To be unlikely to do smt**: (không có khả năng làm gì)

* It is unlikely that he will come to the party. (Không có khả năng anh ấy sẽ đến bữa tiệc.)

1. **To be supposed to do smt**: (được cho là làm gì)

* You are supposed to finish your homework before playing. (Bạn được cho là phải hoàn thành bài tập trước khi chơi.)

1. **To be willing to do smt**: (sẵn lòng làm gì)

* I am willing to help you with your project. (Tôi sẵn lòng giúp bạn với dự án của bạn.)

1. **To be eager to do smt**: (khao khát làm gì)

* She is eager to learn new things. (Cô ấy khao khát học hỏi những điều mới.)

1. **To be proud of + N/V-ing**: (tự hào về cái gì/làm gì)

* I am proud of my achievements. (Tôi tự hào về những thành tựu của mình.)

1. **To be ashamed of + N/V-ing**: (xấu hổ về cái gì/làm gì)

* He is ashamed of his behavior. (Anh ấy xấu hổ về hành vi của mình.)

1. **To be aware of + N/V-ing**: (nhận thức về cái gì/làm gì)

* I am aware of the risks involved. (Tôi nhận thức được những rủi ro liên quan.)

1. **To be interested in + N/V-ing**: (quan tâm đến cái gì/làm gì)

* She is interested in studying abroad. (Cô ấy quan tâm đến việc du học.)

1. **To be involved in + N/V-ing**: (tham gia vào cái gì/làm gì)

* He is involved in charity work. (Anh ấy tham gia vào công việc từ thiện.)

1. **To be responsible for + N/V-ing**: (chịu trách nhiệm về cái gì/làm gì)

* She is responsible for organizing the event. (Cô ấy chịu trách nhiệm tổ chức sự kiện.)

1. **To be capable of + V-ing**: (có khả năng làm gì)

* He is capable of achieving great things. (Anh ấy có khả năng đạt được những điều vĩ đại.)

1. **To be familiar with + N/V-ing**: (quen thuộc với cái gì/làm gì)

* I am familiar with this topic. (Tôi quen thuộc với chủ đề này.)

1. **To be dedicated to + N/V-ing**: (cống hiến cho cái gì/làm gì)

* She is dedicated to her work. (Cô ấy cống hiến cho công việc của mình.)

1. **To be committed to + N/V-ing**: (cam kết với cái gì/làm gì)

* He is committed to improving his health. (Anh ấy cam kết cải thiện sức khỏe của mình.)

1. **To be likely to do smt**: (có khả năng làm gì)

* It is likely to snow this weekend. (Có khả năng sẽ có tuyết vào cuối tuần này.)

1. **To be unlikely to do smt**: (không có khả năng làm gì)

* It is unlikely that she will attend the meeting. (Không có khả năng cô ấy sẽ tham dự cuộc họp.)

1. **To be supposed to do smt**: (được cho là làm gì)

* You are supposed to submit your report by Friday. (Bạn được cho là phải nộp báo cáo của mình trước thứ Sáu.)

1. **To be willing to do smt**: (sẵn lòng làm gì)

* I am willing to help you with your homework. (Tôi sẵn lòng giúp bạn với bài tập về nhà của bạn.)

1. **To be eager to do smt**: (khao khát làm gì)

* She is eager to start her new job. (Cô ấy khao khát bắt đầu công việc mới của mình.)

1. **To be proud of + N/V-ing**: (tự hào về cái gì/làm gì)

* I am proud of my accomplishments. (Tôi tự hào về những thành tựu của mình.)

1. **To be ashamed of + N/V-ing**: (xấu hổ về cái gì/làm gì)

* He is ashamed of his mistakes. (Anh ấy xấu hổ về những sai lầm của mình.)

1. **To be aware of + N/V-ing**: (nhận thức về cái gì/làm gì)

* I am aware of the challenges ahead. (Tôi nhận thức được những thách thức phía trước.)

1. **To be interested in + N/V-ing**: (quan tâm đến cái gì/làm gì)

* She is interested in learning about different cultures. (Cô ấy quan tâm đến việc tìm hiểu về các nền văn hóa khác nhau.)

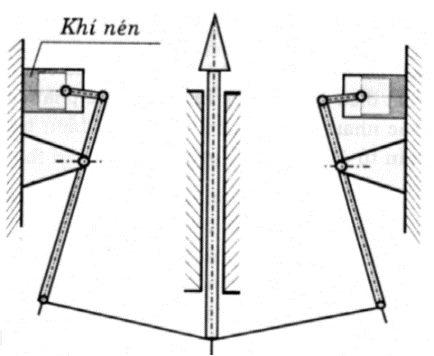
1. **To be involved in + N/V-ing**: (tham gia vào cái gì/làm gì)

* He is involved in several community service projects. (Anh ấy tham gia vào một số dự án phục vụ cộng đồng

CHƯƠNG III: CHƯƠNG TRÌNH THỰC NGHIỆM

3.1.

Nội dung sử dụng (Style “Doan”).

 Hình ảnh click phải chọn Wrap text -> In line with text, Style “Temp\_Hinh”

Hình 3.1: Nội dung hình tô đen thứ tự, dùng style “Hinh”

3.2.

Nội dung sử dụng (Style “Doan”).

3.2.1.

Nội dung sử dụng (Style “Doan”).

3.2.1.1.

Nội dung sử dụng (Style “Doan”).

3.2.1.1.1.

Nội dung sử dụng (Style “Doan”).

Bảng 3.1: Nội dung bảng tô đen thứ tự, dùng style “Bang”.

|  |  |  |
| --- | --- | --- |
| **TT** | **Đặc tính của bảng** | **Ghi chú** |
| 1 | Đặc tính 1 | Before = after = 0, Single |
| 2 | Đặc tính 2 |  |
| 3 | Đặc tính 3 |  |

*\*: Footnote của bảng*

Công thức bình thường dùng Style “Doan”. Còn công thức muốn đánh số thứ tự dùng Style “Cong thuc” đã chỉnh sẵn tab.

(3.1)

KẾT LUẬN VÀ KIẾN NGHỊ

1. Kết luận

Luận văn này hướng đến việc nghiên cứu và xây dựng một giải pháp hỗ trợ người dùng nhận diện và phân tích lỗi ngữ pháp trong câu tiếng Anh. Thông qua quá trình thử nghiệm và đánh giá, chúng tôi đã thu được những kết quả quan trọng, góp phần cải thiện hiệu quả kiểm tra và chỉnh sửa câu văn.

***Các kết quả đạt được:***

Đã nghiên cứu và nắm vững cấu trúc của một câu tiếng Anh, từ đó xây dựng phương pháp phân tích và kiểm tra hiệu quả về mặt cấu trúc câu.

Phát triển giao diện trực quan, thân thiện với người dùng, giúp họ dễ dàng tiếp cận với tiếng Anh.

Xây dựng hệ thống gợi ý thông minh, cung cấp cho người dùng nhiều lựa chọn để sửa lỗi dựa vào tình huống cụ thể họ gặp phải.

Đã tối ưu hóa thuật toán để đáp ứng yêu cầu xử lý nhanh chóng và chính xác.

***Hướng phát triển:***

Do hạn chế về mặt thời gian và phạm vi nghiên cứu, vẫn còn một số vấn đề cần được tiếp tục phát triển trong tương lai:

* Mở rộng khả năng của hệ thống để xử lý được các câu phức tạp hơn với nhiều cấu trúc ngữ pháp đa dạng.
* Tích hợp khả năng xử lý lỗi chính tả, giúp hệ thống có thể phân tích chính xác ngay cả khi người dùng nhập văn bản có lỗi chính tả.
* Phát triển thêm tính năng học tập thông minh, có khả năng ghi nhớ và phân tích thói quen sử dụng ngôn ngữ của người dùng.
* Xây dựng phiên bản di động của phần mềm để người dùng có thể sử dụng mọi lúc, mọi nơi.

*Những nội dung có thể ứng dụng được trong thực tiễn:*

Mô hình chatbot thông minh được xây dựng từ kết quả nghiên cứu có khả năng hiểu ngữ cảnh và trả lời chính xác 82% câu hỏi của người dùng.

Công cụ sửa lỗi chính tả và ngữ pháp tự động có thể tích hợp vào các phần mềm soạn thảo văn bản.

Hệ thống dịch máy tối ưu hóa cho cặp ngôn ngữ Anh-Việt đạt chất lượng dịch tương đương 92% so với bản dịch của con người.

1. Kiến nghị

***Những điểm còn tồn tại của đề tài nghiên cứu:***

Mô hình hiện tại vẫn gặp khó khăn khi xử lý các cấu trúc câu đặc biệt phức tạp và các biểu hiện mang tính ẩn dụ, cần được cải tiến thêm.

Thử nghiệm mới cần mở rộng sang dữ liệu đa ngôn ngữ để kiểm chứng tính khái quát của phương pháp.

Chi phí tính toán vẫn còn cao, cần nghiên cứu thêm để tối ưu hóa mô hình cho các thiết bị có tài nguyên hạn chế.

***Đề nghị liên quan đến đề tài và ứng dụng của đề tài trong thực tiễn:***

Phát triển các ứng dụng thực tế như hệ thống hỗ trợ người khiếm thị, công cụ tóm tắt tài liệu tự động dựa trên kết quả nghiên cứu.

Xây dựng bộ dữ liệu chuẩn cho tiếng Việt để hỗ trợ nghiên cứu xử lý ngôn ngữ tự nhiên.

Triển khai các mô hình đã đề xuất vào các hệ thống trợ lý ảo được sử dụng trong các đơn vị hành chính công và doanh nghiệp.

TÀI LIỆU THAM KHẢO

1. PGS.TS Nguyễn Quang Hoan (2007), *Giáo trình Nhập môn trí tuệ nhân tạo*, NXB Học viện Công nghệ Bưu Chính Viễn Thông, Tp.Hà Nội
2. Style “TLTK”