MỞ ĐẦU

1. Tính cấp thiết của đề tài

Trong nhiều năm trở lại đây, với nhu cầu về hội nhập ngày càng cao giữa các quốc gia cả về kinh tế lẫn văn hóa, yêu cầu về ngoại ngữ, đặc biệt là tiếng Anh, trở thành một vấn đề cấp thiết với mỗi người. Nhưng nhiều người không có đủ thời gian cũng như điều kiện để tham gia các lớp học thêm hoặc các câu lạc bộ để nâng cao trình độ của mình. Chính vì vậy cần có những phần mềm, công cụ để hỗ trợ người học tiếng Anh ở bất cứ nơi đâu, vào bất cứ thời gian nào. Hiện nay đã có nhiều công cụ hướng tới mục đích đó, song mỗi công cụ, phần mềm đều có những hạn chế riêng, đặc biệt là tính thụ động. Người học hầu như chỉ tham gia vào các hoạt động được thiết kế từ trước trên công cụ, ít có sự tương tác hai chiều.

Với sự phát triển của khoa học công nghệ, việc mỗi người sở hữu cho mình một chiếc điện thoại thông minh hiện nay là rất phổ biến. Dựa trên nền tảng điện thoại thông minh mà đã và đang xuất hiện các ứng dụng hỗ trợ học tiếng Anh có sự tương tác cao giữa người và máy; một trong những ứng dụng phổ biến nhất hiện nay có thể kể tới là các ứng dụng dựa trên Chatbot. Tuy nhiên, phần lớn các ứng dụng Chatbot mới tập trung vào phần từ vựng, từ điển, trắc nghiệm... hoặc các ứng dụng luyện kỹ năng nghe, kỹ năng đọc..., có rất ít các ứng dụng có thể giúp người dùng kiểm tra chính tả cũng như cú pháp của câu để chỉ ra lỗi sai cho người đọc và chỉnh sửa nó, mặc dù đây là một bài toán tương đối quan trọng.

Vì vậy, Chatbot đáp ứng được các yêu cầu, chạy trên điện thoại thông minh để hỗ trợ người học tiếng Anh sẽ là một giải pháp có hiệu quả để nâng cao chất lượng học tập tiếng Anh.

Chính vì lý do đó, tác giả đã lựa chọn đề tài: *"Nghiên cứu mô hình PCFGs và ngôn ngữ AIML trong xây dựng chatbot hỗ trợ học tiếng Anh"*.

2. Mục tiêu nghiên cứu

Nghiên cứu cơ sở lý thuyết nền tảng của bài toán kiểm tra chính tả và cú pháp của câu trong tiếng Anh; ứng dụng cài đặt, đánh giá giải thuật và xây dựng một ứng dụng hỗ trợ các tính năng như kiểm tra chính tả, ngữ pháp, cú pháp thông qua hội thoại giữa người dùng và máy trên nền tảng Android.

3. Đối tượng và phạm vi nghiên cứu của đề tài

Đối tượng nghiên cứu

- Khái quát về trí tuệ nhân tạo

- Mô hình PCFGs, ứng dụng xây dựng cây cú pháp

- Ngôn ngữ AIML và kỹ thuật xây dựng chatbot

Phạm vi nghiên cứu

Chatbot trên điện thoại thông minh sử dụng hệ điều hành Android được xây dựng dựa trên AIML và mô hình PCFGs có khả năng thực hiện hội thoại với người dùng, phát hiện và sửa những lỗi chính tả và cú pháp.

4. Phương pháp nghiên cứu

- Khảo sát, phân tích và hệ thống hóa nội dung các tài liệu khoa học liên quan đến chatbot hỗ trợ học tiếng Anh

- Đối sánh nội dung nghiên cứu của đề tài với các nội dung nghiên cứu đã thực hiện để vừa phát triển áp dụng các kết quả khoa học - công nghệ đã có cho đề tài vừa tìm ra các nội dung mới cần được nghiên cứu và thi hành.

- Thiết kế mô hình và thực nghiệm đánh giá các kỹ thuật, bài toán đã đề xuất để chứng minh tính hiệu quả.

5. Ý nghĩa khoa học, ý nghĩa thực tiễn của đề tài

Ý nghĩa khoa học

- Nghiên cứu, nắm vững về trí tuệ nhân tạo và ngôn ngữ AIML

- Vận dụng trí tuệ nhân tạo để tạo ra sự giao tiếp thân thiện, gần gũi giữa người và máy tính

- Tìm hiểu về chatbot và ứng dụng chatbot để cung cấp thông tin

Ý nghĩa thực tiễn

- Tạo ra được công cụ hỗ trợ học tiếng Anh theo hình thức hội thoại giữa người và máy

- Giúp phát hiện và sửa những lỗi thường gặp về chính tả và cú pháp trong quá trình giao tiếp (viết, nói) bằng tiếng Anh.

- Nâng cao hiệu quả học tiếng Anh.

6. Kết cấu luận văn

- Chương 1: Các vấn đề tổng quan: Giới thiệu tổng quan lý thuyết về trí tuệ nhân tạo, xu hướng phát triển của trí tuệ nhân tạo, lĩnh vực xây dựng chatbot hỗ trợ học tiếng Anh, bài toán phân tích cú pháp, kiểm tra chính tả, ngữ pháp và các vấn đề liên quan.

- Chương 2: Mô hình PCFGs và ngôn ngữ AIML: Nghiên cứu văn phạm phi ngữ cảnh, tính mập mờ trong phân tích cú pháp và đề xuất giải pháp sử dụng văn phạm phi ngữ cảnh hướng thống kê PCFGs; nghiên cứu mã nguồn mở AIML trong xây dựng chatbot.

- Chương 3: Phân tích thiết kế, cài đặt ứng dụng: Trình bày cơ bản về thiết kế của ứng dụng và kết quả đạt được thông qua một số mẫu kiểm thử.

- Kết luận: Trình bày điểm mạnh và hạn chế trong luận văn. Đồng thời nêu ra hướng phát triển tiếp theo trong tương lai.

CHƯƠNG 1: CÁC VẤN ĐỀ TỔNG QUAN

1.1. Chatbot

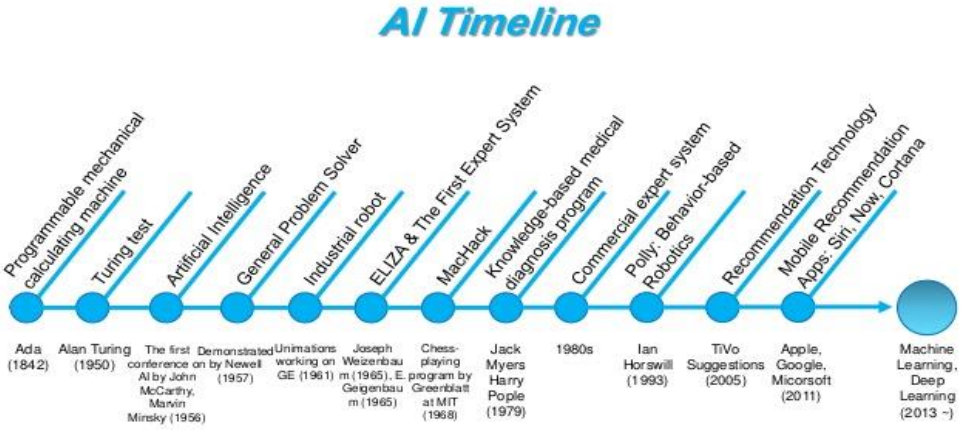
1.1.1. Trí tuệ nhân tạo

1.1.1.1 Định nghĩa

Khái niệm về trí tuệ nhân tạo (Artificial Intelligence - viết tắt là AI) có thể được nhìn nhận theo nhiều cách khác nhau, chưa có định nghĩa nào được thừa nhận chung. Trên thế giới hiện có nhiều định nghĩa về trí tuệ nhân tạo, để đơn giản chúng ta có thể hiểu trí tuệ nhân tạo là một ngành khoa học máy tính. Nó xây dựng trên một nền tảng lý thuyết vững chắc và có thể ứng dụng trong việc tự động hóa các hành vi thông minh của máy tính; giúp máy tính có được những trí tuệ của con người như: biết suy nghĩ và lập luận để giải quyết vấn đề, biết giao tiếp do hiểu ngôn ngữ, tiếng nói, biết học và tự thích nghi[18].

1.1.1.2. Quá trình hình thành và phát triển

Ý tưởng xây dựng một chương trình AI xuất hiện lần đầu vào tháng 10/1950, khi nhà bác học người Anh Alan Turing xem xét vấn đề "liệu máy tính có khả năng suy nghĩ hay không?".



Hình 1.2. Quá trình hình thành và phát triển của trí tuệ nhân tạo

1.1.1.3. Một số ứng dụng

Hiện tại, trí tuệ nhân tạo được ứng dụng trong đời sống theo hai hướng: Dùng máy tính để bắt chước quá trình xử lý của con người và thiết kế những máy tính thông minh độc lập với cách suy nghĩ của con người.

Một số ứng dụng của trí tuệ nhân tạo trong cuộc sống thực tiễn có thể kể đến như: nhận dạng chữ viết, nhận dạng tiếng nói, dịch tự động, tìm kiếm thông tin, khai phá dữ liệu và phát triển tri thức, lái xe tự động, robot[18].

1.1.1.4. Xu thế nghiên cứu và phát triển của trí tuệ nhân tạo hiện đại

- Nhận dạng mẫu
- Xử lý ảnh
- Mạng nơron
- Xử lý ngôn ngữ tự nhiên
- Robot học
- Chatbot...

1.1.1. Chatbot là gì?

Chatbot (có thể được gọi là chatter robot) là một lĩnh vực của trí tuệ nhân tạo. Chatbot là một hệ thống thực hiện sự trao đổi thông tin giữa hai hay nhiều đối tượng theo một quy chuẩn nhất định, quá trình trao đổi thông tin có thể bằng ngôn ngữ nói, ngôn ngữ viết hoặc kí hiệu[2].

Chatbot có thể được phân loại thành 3 loại chính[2]:

- Chatbot giữa người với người
- Chatbot giữa máy với máy
- Chatbot giữa người và máy

1.1.2. Chatbot hỗ trợ học tiếng Anh

1.1.2.1. Miki

Miki là một chatbot trên Facebook, được hoạt động sau khi Facebook chính thức hỗ trợ một nền tảng dành cho bot trên Messenger. Các tính năng học tính Anh được hỗ trợ trên Miki:

- Tra từ điển Anh Việt
- Tra câu song ngữ Anh Việt
- Dịch oạn văn

1.1.2.2. Poli Bot

Poli là một chatbot chuyên dạy thành ngữ tiếng Anh, với một số tính năng như sau:

- Cung cấp các thành ngữ tiếng Anh
- Xem định nghĩa
- Xem các ví dụ về cách dùng

1.1.2.3. Sally Bot

Các tính năng của Sally:

- Học cụm động từ mới
- Định nghĩa cụm từ đã cho
- Đưa ví dụ liên quan đến cụm từ đã cho
- Đưa cụm từ đã cho áp dụng vào đoạn hội thoại

1.1.2.4. Andy English

Các tính năng của Andy English:

- Hội thoại bằng tiếng Anh, thảo luận về các chủ đề khác nhau
- Học ngữ pháp
- Học thêm từ mới để mở rộng vốn từ

1.1.2.5. Acobot

Acobot là một ứng dụng hỗ trợ học tiếng Anh với các tính năng giúp người sử dụng luyện các kỹ năng đọc, viết, nghe, nói, đàm thoại, phát âm, dịch thuật, ngữ pháp và từ vựng.

Qua nghiên cứu các chatbot trên, có thể thấy rằng, hầu như tất cả các chatbot đều tập trung vào việc tập trung vào phần luyện từ vựng, ứng dụng từ điển, trắc nghiệm... hoặc các ứng dụng luyện kỹ năng nghe, kỹ năng đọc; gần như chưa có ứng dụng nào hỗ trợ người sử dụng trong việc kiểm tra chính tả, ngữ pháp. Đó cũng chính là lý do chính để tác giả lựa chọn đề tài này.

1.2. Ngữ pháp tiếng Anh

1.2.1. Các khái niệm cơ bản

Ngữ pháp

Ngữ pháp là quy tắc chủ yếu trong cấu trúc ngôn ngữ. Ngữ pháp, theo cách hiểu của hầu hết các nhà ngôn ngữ học hiện đại bao gồm ngữ âm, âm học, hình thái ngôn ngữ, cú pháp, ngữ nghĩa. Kiểm tra ngữ pháp là quá trình kiểm tra một văn bản có phù hợp với ngữ pháp của ngôn ngữ đó hay không.

Cú pháp

Cú pháp là một phần trong ngữ pháp. Cú pháp bao gồm tập các luật, nguyên tắc và các quá trình biến đổi để ta có thể xây dựng cấu trúc của một câu trong một ngôn ngữ theo một thứ tự nhất định.

Các lớp từ (nhãn từ) trong tiếng Anh

Các thành phần ngữ pháp có thể được chia thành 2 mảng lớn: đóng và mở. Có 4 lớp từ mở chính: Danh từ (nouns), động từ (verbs), tính từ (adjectives) và trạng từ (adverbs). Tuy nhiên, điều này đúng với tiếng Anh nhưng không phải với tất cả các ngôn ngữ, nhiều ngôn ngữ không có tính từ.

Các lớp từ đóng khác nhau giữa các ngôn ngữ khác nhau hơn so với các lớp mở. Dưới đây là tổng quát một vài lớp từ đóng quan trọng trong tiếng Anh:

- Giới từ (Prepositions): on, under, over, near, by, from, to, with...
- Mạo từ (Determiners): a, an, the...
- Đại từ nhân xưng (Pronouns): I, she, he, who...
- Liên từ (Conjunctions): and, but, or, as, if, when...
- Trợ động từ (Auxiliary verbs): can, may, should, are...
- Particles: up, down, on, off, in, out, at, by...
- Số đếm (numerals): one, two, three...

1.2.2. Phân loại lỗi

- Lỗi chính tả (Spelling errors)
- Lỗi ngữ pháp (Grammar errors)
- Lỗi phong cách dùng từ (Style errors)

1.2.3. Một số lỗi ngữ pháp trong tiếng Anh

- Lỗi chia động từ (Subject-Verb Agreement)[10]
- Lỗi dùng mạo từ không xác định a/an[10]
- Câu hỏi đuôi (Tag questions)[10]
- Những lỗi khác

1.3. Tổng quan bài toán kiểm tra ngữ pháp tiếng Anh

Để giải quyết bài toán kiểm tra ngữ pháp tiếng Anh, chúng ta cần phải thực hiện 2 nhiệm vụ:

- Phân tích cú pháp
- Kiểm tra ngữ pháp

1.3.1. Phân tích cú pháp

1.3.1.1. Xử lý ngôn ngữ tự nhiên và các vấn đề chính

Hình 1.8. Mô hình xử lý ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên có rất nhiều ứng dụng trong thực tế, có thể kể ra ở đây một vài ứng dụng của xử lý ngôn ngữ tự nhiên như là dịch máy (machine translation), tìm kiếm thông tin (information retrieval), trích chọn thông tin (information retrieval) hay như là nhận dạng tiếng nói (speech recognition)[6].

1.3.1.2. Phân tích cú pháp

Phân tích cú pháp (parsing analysis hay syntatic analysis) là quá trình phân tích một chuỗi từ tố (chuỗi từ tố này là kết quả của quá trình phân tích từ tố, thông thường đối với xử lý ngôn ngữ là các từ), nhằm đưa ra các cấu trúc ngữ pháp của chuỗi từ đó dựa vào một văn phạm nào đó. Thông thường cấu trúc ngữ pháp được chọn ở đây thường là dạng cây, bởi thông qua dạng này sự phụ thuộc của các thành phần là trực quan[6].

Nói cách khác, phân tích cú pháp là quá trình dựa vào văn phạm để xây dựng một cây cú pháp.

Trong tiếng Anh, phân tích cú pháp cụ thể là phân tích một câu và xây dựng một cây cú pháp dựa trên một văn phạm, văn phạm đó thường là dựa trên tập luật ngữ pháp trong tiếng Anh. Ta sẽ kiểm tra câu hoặc văn bản có thỏa mãn các luật ngữ pháp trong tiếng Anh hay không. Nếu một câu không thể xây dựng thành công cây cú pháp, có nghĩa câu văn đó là lỗi.

1.3.1.3. Vai trò của phân tích cú pháp trong xử lý ngôn ngữ tự nhiên

Có thể nói phân tích cú pháp là bài toán cơ sở, xuất hiện rất nhiều trong các ứng dụng của xử lý ngôn ngữ tự nhiên. Ứng dụng đầu tiên ta có thể thấy ngay đó là áp dụng phân tích cú pháp trong kiểm tra lỗi ngữ pháp. Đối với việc kiểm tra lỗi ngữ pháp ta cần thực hiện việc phân tích cú pháp câu đầu vào, xem cấu trúc có đúng không?

1.3.1.4. Các hướng tiếp cận

Để tiếp cận bài toán, có 2 hướng chính: Phương pháp phân tích từ trên xuống (Top - Down Parsing) và phương pháp phân tích từ dưới lên (Bottom - Up Parsing). Những thuật ngữ này là dựa vào thứ tự xây dựng các nút trong cây phân tích cú pháp. Phương pháp Top - Down là bắt đầu xây dựng từ gốc tiến hành hướng xuống các nút lá, còn phương pháp Bottom - Up là tiếp cận từ các lá tiến về gốc.

1.3.2. Bài toán kiểm tra ngữ pháp tiếng Anh

- Kiểm tra dựa vào cú pháp (Syntax-based checking)
- Kiểm tra dựa vào thống kê (Statistics-based checking)
- Kiểm tra dựa vào luật (Rule-based checking)

1.4. Kết luận chương

CHƯƠNG 2: MÔ HÌNH PCFGs VÀ NGÔN NGỮ AIML

2.1. Mô hình PCFGs

2.1.1. Văn phạm phi ngữ cảnh

Các khái niệm cơ bản

Một văn phạm phi ngữ cảnh (CFG) là một tập 4 thành phần chính G = (N, ∑, R, S), trong đó:

- N là tập chứa hữu hạn các phần tử được gọi là phần tử không kết thúc

- ∑ là tập chứa hữu hạn các phần tử được gọi là phần tử kết thúc

- R là tập các luật ngữ pháp có dạng $X \rightarrow Y_1Y_2...Y_n$, $X \in N$, $n \geq 0$, $Y_i \in (N \cup \sum)$ với i = 1...n.

- S là một trong những phần tử ∈ N được gọi là ký tự bắt đầu.

Dẫn xuất trái (Left-most Derivations)

Cho một văn phạm phi ngữ cảnh G, một dẫn xuất trái là một chuỗi các xâu $s_1...s_n$, trong đó:

$s_1 = S$, cụ thể s_1 chứa một thành phần đơn là ký tự bắt đầu.

$s_n \in \sum^*$, s_n được tạo thành từ các phần tử kết thúc, cụ thể là các thành phần thuộc tập ∑ (viết $\sum^*$ để chỉ tập tất cả các xâu có thể được tạo thành từ các từ trong tập ∑).

Mỗi s_i (i = 2...n) là dẫn xuất từ s_{i-1} bằng cách lấy cách lấy các phần tử không kết thúc gần nhất bên trái X và thay thế chúng bằng các α trong đó α là tập luật phải được tạo ta từ X trong tập R, nói cách khác $X \rightarrow \alpha$.

2.1.2. Tính mập mờ trong phân tích cú pháp

Như đã đề cập ở trên, một xâu S có thể có nhiều hơn một dẫn xuất có thể thực hiện, trong trường hợp này, ta nói xâu s là mập mờ.

Tính mập mờ là một vấn đề rắc rối trong ngôn ngữ tự nhiên. Khi các nhà nghiên cứu lần đầu xây dựng một ngữ pháp đủ lớn phù hợp cho các ngôn ngữ như tiếng Anh, họ phát hiện rằng các câu văn thường có một lượng lớn cây cú pháp có thể xây dựng.

2.1.3. Văn phạm phi ngữ cảnh hướng thống kê PCFGs

2.1.3.1. Các khái niệm cơ bản

Cho G là một văn phạm phi ngữ cảnh, ta có cách khái niệm sau:

- T_G là tập hợp tất cả các cây cú pháp có thể xây dựng được trong G. Khi G rỗng ta có thể viết đơn giản tập hợp này là T.

- Với bất kỳ cây cú pháp t ∈ T_G, ta có yield(t) để mô tả xâu s ∈ ∑*, s cũng là xâu gồm chuỗi các từ được tạo ra từ t.

- Ta có câu s ∈ ∑*, ta định nghĩa T_G(s):

$$T_G(s) = \{t: t \in T_G, \text{yield}(t) = s\} \quad (2.1)$$

Nói cách khác, T_G(s) là tập tất cả các cây cú pháp tạo thành xâu s.

- Ta nói một câu s là mập mờ nếu: $|T_G(s)| > 1$.

- Ta nói một câu s là đúng cú pháp nếu: $|T_G(s)| > 0$.

2.1.3.2. PCFGs (Probabilistic Context-Free Grammars)

Một văn phạm phi ngữ cảnh hướng thống kê (PCFGs - Probabilistic Context-Free Grammars), còn được biết đến với tên SCFG (Stochastic Context-Free Grammar) được đề xuất lần đầu bởi Booth (1969), được định nghĩa như sau:

Một PCFGs bao gồm:

- Một văn phạm phi ngữ cảnh G = (N,∑,S,R).

- Một tham số q(A →B) là xác suất xảy ra của luật A → B trong một dẫn xuất trái. Với X ∈ N, ta có:

$$\sum_{A \to B \in R, A = X} q(A \to B) = 1 \quad (2.3)$$

Hơn nữa q(A→B) ≥ 0 với mọi A → B ∈ R.

Cho một cây cú pháp t ∈ T_G chứa các luật $A_1 \to B_1$, $A_2 \to B_2$... $A_n \to B_n$, xác suất của t khi sử dụng PCFGs là:

$$p(t) = \prod_{i=1}^{n} q(A_i \to B_i) \quad (2.4)$$

Một cách đơn giản, thực hiện xây dựng cây cú pháp dựa trên PCFGs theo các bước sau đây:

1. Khởi tạo s_1 = S, i = 1.

2. Trong khi s_i chứa ít nhất một kí tự chưa kết thúc:

- Tìm luật trái trong s_i, gọi là X.

- Chọn một luật có dạng X → A từ tập luật cùng q(X →A).

- Tạo s_i+1 bằng cách thay thế X trong s_i bằng A.

- Đặt i = i + 1 và lặp lại quá trình.

2.1.3.3. Xây dựng PCFGs từ kho dữ liệu (Corpus)

Giả thiết rằng ta đã có một tập dữ liệu huấn luyện gồm các cây cú pháp t_1, t_2... t_m. Khi đó yield(t_i) chỉ câu được tạo ra từ cây cú pháp thứ i, cũng là câu thứ i trong kho dữ liệu.

Mỗi cây cú pháp t_i là gồm một tập các luật phi ngữ cảnh, giả sử tất cả các cây cú pháp trong kho dữ liệu đều có gốc là S, khi đó ta định nghĩa một PCFGs(N,∑,S,R,q) như sau:

- N là tập các phần tử không kết thúc trong các t_1, t_2,… t_m.

- ∑ là tập các từ trong các cây t_1, t_2,… t_m.

- S là ký hiệu bắt đầu.

- R là tập luật bao gồm tất cả các luật có dạng A→B trong t_1, t_2,… t_m.

- q là thông số xác suất của từng luật trong tập R, được tính theo công thức:

$$q(A \to B) = \frac{Count(A \to B)}{Count(A)} \tag{2.5}$$

Trong đó: Count(A→B) là số lần xuất hiện luật A→B trong kho dữ liệu, Count(A) là số lần xuất hiện của các luật có dạng A→X, X ∈ N trong kho dữ liệu.

2.1.3.4. Xây dựng cây cú pháp với PCFGs

Luận văn sẽ tập trung tìm hiểu sử dụng thuật toán CKY để giải quyết bài toán xây dựng cây cú pháp với PCFGs. Thuật toán là dựa trên thuật toán CKY (Cocke-Kasami-Younger) hướng xác suất, được đưa ra lần đầu bởi Ney năm 1991.

Ngữ pháp chuẩn Chomsky (CNF)

Định nghĩa: Một văn phạm phi ngữ cảnh G = (N,∑,R,S) được gọi thỏa mãn chuẩn Chomsky nếu mỗi luật A→B ∈ R đều có một trong hai dạng sau:

- $X \to Y_1Y_2$, X ∈ N, Y_1∈ N, Y_2∈ N.

- X → Y, X ∈ N, Y ∈ ∑

Chuyển đổi về dạng CNF

Yêu cầu đặt ra là chuyển đổi một ngữ pháp PCFGs không theo chuẩn CNF về dạng CNF.

Ta có các trường hợp sau:

- Luật $X \in R$ có dạng $X \rightarrow Y_1\ Y_2\ Y_3$

Ta tiến hành biến đổi

$$Y_1\_Y_2 \rightarrow Y_1\ Y_2$$

$$X \rightarrow Y_1\_Y_2\ Y_3$$

Như vậy ta có thể biến đổi một luật từ không thuộc CNF về dạng CNF.

Thuật toán CKY xây dựng cây cú pháp với văn phạm PCFGs

Trong phần này ta sẽ trình bày một thuật toán để phân tích cây cú pháp với văn phạm PCFGs có chuẩn CNF.

Dữ liệu đầu vào là một PCFGs $G = (N,\sum,S,Q,q)$ với chuẩn CNF, và một câu $s = x_1x_2\ldots x_n$ với x_i là từ thứ i trong câu.

Đầu ta của thuật toán là kết quả:

$$\arg\max_{t \in T_G(s)} p(t)$$

Thuật toán CKY là một thuật toán quy hoạch động. Ý tưởng chính của thuật toán như sau:

- Cho một câu có sẵn $x_1\ldots x_n$, ta định nghĩa T(i,j,X) cho bất kỳ $X \in N$ và (i,j) thỏa mãn $1 \leq i \leq j \leq n$ là tập gồm tất cả cây cú pháp cho các từ $x_i\ldots x_j$ có gốc là X.

- Ta định nghĩa

$$\pi(i,j,X) = \max_{t \in T(i,j,X)} p(t) \qquad (2.6)$$

$$(\pi(i,j,X) = 0 \ if\ T(i,j,X) = \varnothing)$$

Do đó $\pi(i,j,X)$ là điểm số cao nhất trong tất cả các cây cú pháp tạo thành từ các từ $x_i\ldots x_j$ và có X là gốc. Điểm số đó của cây t có được thông qua các điểm số của các luật mà nó chứa.

Ví dụ $A_1 \rightarrow B_1$, $A_2 \rightarrow B_2$,... $A_n \rightarrow B_n$, theo công thức (2.4) ta có

$$p(t) = \prod_{i=1}^{n} q(A_i \rightarrow B_i)$$

Do vậy

$$\pi(1,n,S) = \max_{t \in T_G(s)} p(t) \tag{2.7}$$

Đặc biệt, trong thuật toán CKY này, ta có thể sử dụng π như một hàm đệ quy để thực hiện các bước tính toán.

Đây là một thuật toán tiếp cận theo hướng "Bottom - Up", ta sẽ tiến hành thực hiện tính toán π(i,j,S) tại j = i đầu tiên, sau đó sẽ tiếp tục với j = i+1,…

Ta có, với mọi i = 1…n, X ∈ N

$$\pi(i,i,X) = \begin{cases} q(X \to x_i) \; if \; X \to x_i \in R \\ 0 \qquad\qquad otherwise \end{cases} \tag{2.8}$$

Lại có với $1 \le i \le j \le n$, X ∈ N

$$\pi(i,j,X) = \max_{\substack{X \to Y\ Z \in R \\ s \in \{i \ldots (j-1)\}}} (q(X \to Y\ Z) \times \pi(i,s,Y) \times \pi(s+1,j,Z)) \tag{2.9}$$

- Thuật toán CKY

Đầu vào: câu $s = x_1 \ldots x_n$, văn phạm phi ngữ cảnh hướng thống kê PCFGs G = (N,∑,S,R,q)

Khởi tạo:

For all i ∈ {1…n}, for all X ∈ N,

$$\pi(i,i,X) = \begin{cases} q(X \to x_i) \; if \; X \to x_i \in R \\ 0 \qquad\qquad otherwise \end{cases}$$

Thuật toán:

- For k = 1…(n-1)

- For i = 1…(n-1)

  + Đặt j = i +k

  + For all X ∈ N

$$\pi(i,j,X) = \max_{\substack{X \to Y\ Z \in R \\ s \in \{i \ldots (j-1)\}}} (q(X \to Y\ Z) \times \pi(i,s,Y) \times \pi(s+1,j,Z))$$

và

$$bp(i,j,X) = \arg \max_{\substack{X \to Y\ Z \in R \\ s \in \{i \ldots (j-1)\}}} (q(X \to Y\ Z) \times \pi(i,s,Y) \times \pi(s+1,j,Z))$$

Đầu ra: $\pi(1,n,S) = \max_{t \in T_G(s)} p(t)$

2.1.3.5. Ưu điểm và hạn chế của hướng tiếp cận PCFGs

Ưu điểm

- PCFGs đưa ra hướng tiếp cận xây dựng một cây cú pháp có độ khả thi.

- Có thể loại bỏ những ngữ pháp không hợp lý và các lỗi ngữ pháp vì khi đó cây cú pháp có thông số, xác suất thấp.

- Giải quyết được vấn đề mập mờ do PCFGs sử dụng xác suất để lựa chọn cây cú pháp phù hợp nhất.

- Có thể phát triển mở rộng, số văn bản được phân tích càng nhiều, PCFGs càng thêm chính xác do xác suất từng luật cũng được điều chỉnh.

- PCFGs rất đơn giản và mô hình xác suất đơn giản đối với cấu trúc cây, mô hình toán học đơn giản, thuật toán không quá phức tạp[6].

Nhược điểm

- PCFGs quan tâm đến cú pháp nhiều hơn là ngữ nghĩa, do vậy đôi khi cây cú pháp được chọn phù hợp về cú pháp nhưng lại không phù hợp về nghĩa.

- Do được tính toán thông qua từng cây con trong cây cú pháp, PCFGs có xu hướng tốt hơn với câu ngắn và giảm dần với các câu dài.

- Khi 2 cây cú pháp được tạo ra có cùng xác suất, PCFGs sẽ chọn cây cú pháp đầu tiên. Điều này đôi khi không chính xác.

- Với số lượng luật trong PCFGs tăng lên, công sức bỏ ra khi xây dựng cây cú pháp cũng tăng lên nhanh chóng.

2.2. Ngôn ngữ AIML

2.2.1. AIML là gì?

AIML (Artificial Intelligence Markup Language) là ngôn ngữ lập trình trí tuệ nhân tạo bắt nguồn từ XML (Extensible Mark-up Language), được sử dụng để phát triển các ứng dụng trí tuệ nhân tạo.

2.2.2. Các Category và đặc tính của AIML

Category là một đơn vị cơ bản trong ngôn ngữ AIML, nó bao gồm một câu hỏi đầu vào, một câu trả lời làm đầu ra và một ngữ cảnh nhất định. Câu hỏi được

đặt trong các thẻ <pattern> và nội dung trong thẻ <template> là câu trả lời tương ứng. Ngữ cảnh của một category được dựa vào 2 loại thẻ là <that> và <topic>.

2.2.3. Một số thẻ thông dụng trong AIML

<aiml>	Định nghĩa bắt đầu và kết thúc của một tài liệu AIML
<category>	Định nghĩa một đơn vị tri thức gồm câu hỏi và câu trả lời
<pattern>	Định nghĩa một mẫu dữ liệu có thể so khớp với đầu vào từ người dùng
<template>	Định nghĩa phản hồi đến người dùng ứng với đầu vào phù hợp pattern
<star>	Được sử dụng để khớp với kí tự * trong dữ liệu từ thẻ <pattern>
<srai>	Được sử dụng để cho phép định nghĩa một phản hồi cho nhiều đầu vào có mục đích tương tự nhau
<random>	Được sử dụng để lấy ngẫu nhiên một phản hồi trong một tập các phản hồi được định nghĩa sẵn
<get>, <set>	Được sử dụng để làm việc với các biến trong AIML, các biến có thể được truy xuất dữ liệu trong qua các thẻ này
<that>	Được sử dụng để đưa ra các phản hồi tùy thuộc theo ngữ cảnh
<topic>	Được sử dụng để lưu trữ ngữ cảnh phục vụ cho việc các đoạn hội thoại sau có thể được diễn ra dựa trên ngữ cảnh đó
<think>	Được dùng để lưu trữ các biến mà không cần thông báo cho người dùng
<condition>	Được dùng để điều chỉnh các phản hồi tùy thuộc vào từng điều kiện rẽ nhánh thích hợp

2.2.4. ProgramAB

Là một chương trình mã nguồn mở được phát triển bởi Richard Wallace và cài đặt trên ngôn ngữ Java, hoạt động với AIML 2.0 và hiện nay đang là nền tảng được ưu tiên với những tính năng mới[12].

Chương trình được xây dựng với cấu trúc để cho phép các lập trình viên dễ dàng mở rộng AIML với những thẻ có thể tự định nghĩa. ProgramAB có thể được sử dụng trong nhiều cách:

- Chạy ProgramAB để giao tiếp với một chatbot.

- Phân tích các file logs và phát triển nội dung của các bot.

- Sử dụng ProgramAB như một thư viện để phát triển ứng dụng trên Java và phát triển các tính năng khác.

- Lập trình các ứng dụng riêng với những thẻ AIML tự định nghĩa.

- Điều chỉnh và xây dựng ProgramAB theo yêu cầu cần thiết.

Với mỗi chương trình sử dụng ProgramAB, ta có thể tự điều chỉnh nội dung các cuộc hội thoại cũng như phát triển thêm các cuộc hội thoại thông qua điều chỉnh bots. Cụ thể ta có thể thêm các dữ liệu AIML tự định nghĩa vào các bot hoặc tích hợp nhiều bot vào tùy yêu cầu. Hiện nay, trong ProgramAB mặc định có sẵn dữ liệu của bot như Alice, tuy nhiên ta có thể tìm hiểu thêm một số bot tương tự như Anna, Charlie, Super…

2.3. Kết luận chương

CHƯƠNG 3: PHÂN TÍCH THIẾT KẾ, CÀI ĐẶT ỨNG DỤNG

3.1. Phân tích thiết kế

3.1.1. Xác định yêu cầu

3.1.1.1. Chức năng chính

Hội thoại giữa người và máy

- Người dùng có thể giao tiếp với máy thông qua các đoạn hội thoại, với mỗi câu hội thoại của người dùng, máy sẽ tự động tìm câu trả lời thích hợp và đưa ra cho người dùng.

- Trong dữ liệu của máy sẽ có nhiều chủ đề, người dùng có thể chọn chủ đề để nói chuyện hoặc có thể thay đổi chủ đề bất kỳ khi nói chuyện.

Kiểm tra chính tả, ngữ pháp

Ứng dụng tập trung kiểm tra chính tả và ngữ pháp của các dữ liệu từ người dùng. Tuy nhiên, chức năng này mới chỉ hỗ trợ kiểm tra chính tả và kiểm tra ngữ pháp nghiêng về chia động từ trong câu, không bao gồm kiểm tra cú pháp câu.

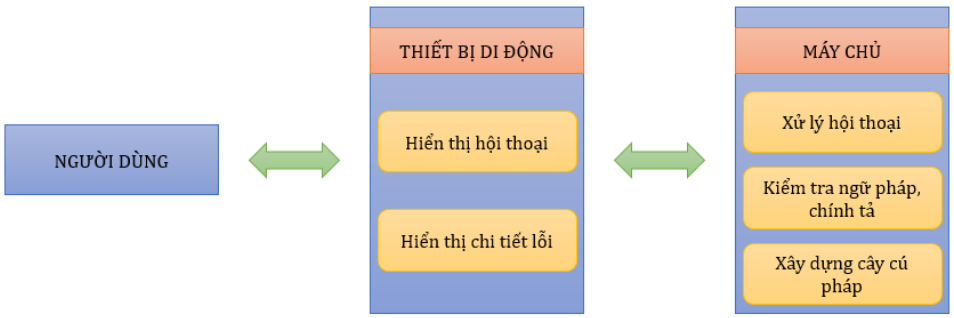
Chi tiết chức năng:

- Người dùng có thể tiến hành đưa dữ liệu vào ứng dụng để ứng dụng kiểm tra lỗi chính tả và ngữ pháp.

- Người dùng có thể chỉ xem lỗi của câu hoặc có thể xem chi tiết cách sửa lỗi từ ứng dụng.

- Dữ liệu đầu vào và đầu ra từ ứng dụng vẫn theo dạng hội thoại.

3.1.1.2. Mô hình hệ thống



Hình 3.1. Mô hình hệ thống AI English

Máy chủ (Server)

- Phụ trách xử lý dữ liệu từ client gửi lên và xử lý, trả về kết quả cho client.

- Các chức năng chính thực hiện trên server: Nhận hội thoại từ người dùng, tìm câu trả lời thích hợp và trả về cho người dùng, kiểm tra chính tả, ngữ pháp của câu người dùng gửi lên, từ câu người dùng gửi lên, xây dựng cây cú pháp và trả về cho người dùng.

Thiết bị di động (Device)

- Thiết bị di động thực hiện nhiệm vụ chính là hiển thị nội dung hội thoại, nhận yêu cầu người dùng gửi đi server và chờ xử lý. Các nội dung như lỗi chi tiết hay sửa lỗi, cây cú pháp cũng được hiển thị trên thiết bị di động.

- Các thiết bị di dộng cần sử dụng nền tảng Android 4.2 trở lên để cài đặt và sử dụng ứng dụng.

3.1.1.3. Chức năng người dùng

Người sử dụng sẽ có những chức năng sau:

- Đăng ký thông tin cá nhân

- Hội thoại bằng tin nhắn với chatbot

- Tự động kiểm tra chính tả

- Tự động kiểm tra ngữ pháp

- Dịch các tin nhắn sang tiếng Việt

- Nghe nội dung tin nhắn

- Tra cứu động từ bất quy tắc

- Tra cứu một số câu thông dụng

3.1.2. Xây dựng tập luật dựa trên tập dữ liệu có sẵn

Quá trình xây dựng tập luật CNF trong PCFGs gồm 3 bước chính:

- Xây dựng kho dữ liệu câu tiếng Anh.

- Xử lý các câu trong kho dữ liệu và tạo các luật cơ bản theo chuẩn CNF.

- Xây dựng tập luật theo văn phạm PCFGs từ các luật cơ bản.

3.1.2.1. Xây dựng kho dữ liệu câu tiếng Anh

Nhiệm vụ của phần này là yêu cầu xây dựng một kho dữ liệu gồm nhiều câu tiếng Anh để phục vụ xây dựng các tập luật.

Để thực hiện nhiệm vụ này, ta sử dụng tập dữ liệu của Tatoeba[20], đây là một trang web bao gồm nhiều tập dữ liệu câu bằng nhiều ngôn ngữ khác nhau. Tuy nhiên, tập dữ liệu câu lấy về từ Tatoeba có chứa nhiều ngôn ngữ khác nhau. Ta thực hiện tách dữ liệu các câu tiếng Anh theo thẻ **"eng"** và thu được tập dữ liệu các câu tiếng Anh cần dùng có dạng như sau:

Bằng cách như vậy, ta có được tập dữ liệu gồm 885113 câu tiếng Anh từ tập dữ liệu ban đầu (số lượng câu tiếp tục được tăng lên). Tập dữ liệu này sẽ được sử dụng để xây dựng các tập luật cơ bản ở phần tiếp theo.

3.1.1.2. Xử lý các câu trong kho dữ liệu và tạo các luật cơ bản theo chuẩn CNF

Để thực hiện điều này, ta sử dụng thư viện Stanford-parser[21] để xây dựng một cây cú pháp từ một câu bất kỳ, sau đó từ cây cú pháp có được, ta tách ra từng luật cú pháp và thêm vào tập dữ liệu cú pháp CFG. Trong quá trình tách các luật, ta cố gắng đưa các luật về chuẩn CNF.

3.1.1.3. Xây dựng tập luật theo văn phạm PCFGs

Từ tập luật đã có từ phần 3.1.2.2, ta tiến hành xây dựng tập luật để sử dụng trong văn phạm PCFGs. Để thực hiện điều này, tại mỗi luật ta tiến hành tính xác suất của chúng theo công thức (2.5) đã được đề cập như sau:

$$q(A \to B) = \frac{Count(A \to B)}{Count(A)}$$

3.2. Cài đặt ứng dụng

AI English là một ứng dụng được thiết kế trên các thiết bị di động hệ điều hành Android, giao tiếp với người dùng thông qua hội thoại, tương tự các ứng dụng chatbot. Ứng dụng có 2 chức năng chính: Hội thoại, kiểm tra ngữ pháp (chủ yếu là kiểm tra chính tả và chia động từ), cùng với đó là một số tính năng khác như: tra cứu từ điển, nghe nội dung hội thoại, tra cứu động từ bất quy tắc và các câu thông dụng.

3.2.1. Giao diện chức năng hội thoại (Chatbot)

Giao diện gồm các phần sau đây:

- Thanh nhập dữ liệu để người dùng nhập văn bản, một nút bấm để thực hiện gửi yêu cầu từ người dùng đến ứng dụng.

- Bên phải khung hình hiển thị dữ liệu người dùng nhập vào.

- Bên trái khung hình hiển thị trả lời từ ứng dụng.

- Các nút bấm hỗ trợ người dùng đọc dữ liệu bằng tiếng Anh và dịch dữ liệu sang tiếng Việt.

3.2.2. Giao diện chức năng tra cứu từ điển

Giao diện chức năng này được thực hiện trực tiếp trên các đoạn hội thoại, khi người sử dụng muốn tra cứu nghĩa của một từ nào đó.

3.2.3. Giao diện chức năng kiểm tra chính tả, ngữ pháp

Giao diện chức năng này có một số sự khác biệt với chức năng hội thoại như sau:

- Bên trái khung hình hiển thị xử lý lỗi của câu từ người dùng.

- Phần chữ được in đỏ thể hiện lỗi sai của câu.

- Bên cạnh khung hình có một nút để người dùng có thể xem chi tiết lỗi và sửa lỗi.

Khi người dùng muốn hiển thị lỗi và sửa lỗi, giao diện chi tiết sẽ gồm 2 phần: Câu lỗi và câu sửa lỗi. Phần lỗi sẽ được tô đỏ và phần sửa lỗi sẽ được to xanh để phân biệt.

Ngoài ra, ứng dụng còn hỗ trợ người dùng với một bảng động từ bất quy tắc và danh sách, cách sử dụng các cụm từ thông dụng.

3.3. Đánh giá ứng dụng

Ứng dụng AI English đã cơ bản giải quyết được vấn đề kiểm tra ngữ pháp và phân tích cây cú pháp của câu trong tiếng Anh. Bằng cách sử dụng các đoạn hội thoại và giao diện đơn giản, ứng dụng tạo có thể giao tiếp với người dùng một cách dễ dàng và tiện lợi. Ứng dụng có thể giúp người dùng kiểm tra các câu đơn giản cũng như xem xét cú pháp câu để nâng cao khả năng về cách kết hợp các thành phần câu. Ngoài ra, ứng dụng cũng là một ứng dụng hội thoại có thể giúp người dùng trò chuyện để nâng cao vốn kiến thức.

Tuy nhiên, ứng dụng vẫn còn một số vấn đề như:

- Do kết hợp nhiều mã nguồn mở và phương pháp nên ứng dụng còn hạn chế về mặt xử lý cùng lúc nhiều yêu cầu như cả kiểm tra ngữ pháp, cả kiểm tra cú pháp và tích hợp cùng hội thoại.

- Thời gian xử lý cú pháp và ngữ pháp vẫn tương đối dài, đặc biệt đối với nhưng câu dài, gồm nhiều thành phần.

- Giao diện phần chi tiết cú pháp câu có thể gây khó hiểu với những người dùng không quen thuộc với các thành phần câu cũng như viết tắt của các thành phần câu đó.

- Với tính năng nhận dữ liệu đầu vào là âm thanh, ứng dụng yêu cầu phụ thuộc vào phần cứng điện thoại với chức năng voice tốt và điều kiện sử dụng là môi trường yên tĩnh.

- Đánh giá hiệu quả: với 100 câu tiếng Anh có lỗi là dữ liệu đầu vào thì ứng dụng đã phát hiện ra 81 câu bị lỗi, đạt 81%.

KẾT LUẬN

Đóng góp của luận văn

Trong thực tế, việc kiểm tra ngữ pháp và phân tích cú pháp câu tiếng Anh có thể được áp dụng trong nhiều ứng dụng như chia động từ, kiểm tra ngoại ngữ, học ngữ pháp trong tiếng Anh. Trong quá trình nghiên cứu về xây dựng ứng dụng hỗ trợ kiểm tra ngữ pháp tiếng Anh, luận văn đã đạt được một số nội dung sau:

- Luận văn đã tìm hiểu một số định nghĩa cơ bản về ngữ pháp trong tiếng Anh, một số hướng tiếp cận cơ bản bài toán kiểm tra ngữ pháp.

- Tìm hiểu cách tiếp cận PCFGs và áp dụng thuật toán CKY trong bài toán phân tích cú pháp câu.

- Giới thiệu một số mã nguồn mở hỗ trợ xây dựng Chatbot và kiểm tra ngữ pháp. Cài đặt ứng dụng đơn giản giải quyết bài toán kiểm tra ngữ pháp với tập luật tự huấn luyện.

Hạn chế của luận văn

Trong quá trình hoàn thành luận văn và ứng dụng, mặc dù đã đạt được một số kết quả nhất định trong bài toán phân tích ngữ pháp và xây dựng cú pháp tiếng Anh, vẫn có những hạn chế nhất định:

- Phương pháp tiếp cận phụ thuộc nhiều vào tập luật, do vậy khi tập luật có kích thương đủ lớn, thời gian xử lý của ứng dụng là lớn. Hơn nữa, độ chính xác của kết quả cũng phụ thuộc vào độ chính xác của tập luật.

- Kết quả của quá trình phân tích cú pháp chỉ có thể đưa ra cây cú pháp nếu văn bản phù hợp, nhưng chưa thể đưa ra lỗi cụ thể đối với nhưng văn bản lỗi.

- Tập dữ liệu hội thoại sử dụng trong Chatbot còn đơn giản.

- Ứng dụng mới chỉ tập trung vào mặt cú pháp, chưa thể hiện được về mặt ngữ nghĩa của câu.

Hướng phát triển

Với sự phong phú và cấp thiết của tiếng Anh, luận văn có nhiều hướng có thể phát triển tiếp tục:

- Tìm hiểu phương pháp tối ưu tập luật hiện tại, mở rộng tập luật, giảm thiểu thời gian với những tập luật gán nhãn, nâng cao hiệu quả phát hiễn lỗi của ứng dụng.

- Bổ sung tập dữ liệu hội thoại cho hệ thống Chatbot, phong phú về chủ đề.

- Phát triển các tình năng hỗ trợ học các ngữ pháp cụ thể như các dạng câu chủ động, bị động hay các thì trong tiếng Anh.

- Tìm hiểu phương pháp xác định lỗi cụ thể với bài toán phân tích cú pháp câu với những câu lỗi và đưa ra gợi ý sửa lỗi.