HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE & ENGINEERING

XỬ LÍ NGÔN NGỮ TỰ NHIÊN (CO3085)

Report topic

Tokenization and Text Classification

Giáo viên hướng dẫn: thầy Võ Thanh Hùng
Lớp: CN01
Sinh viên thực hiện:
Nguyễn Trang Sỹ Lâm-2152715
Nguyễn Hoàng Khôi Nguyên-2152809

Thành phố Hồ Chí Minh, tháng 3/2024

Mục lục

1. Lời mở đầu

Phân tích cú pháp câu (parsing) là một nhiệm vụ quan trọng trong xử lý ngôn ngữ tự nhiên (NLP), nhằm xác định cấu trúc ngữ pháp của một câu. Trong tiếng Nhật, quá trình này đặc biệt phức tạp do tính linh hoạt về thứ tự từ và sự phụ thuộc vào ngữ cảnh. Tuy nhiên, phân tích cú pháp chính xác là cần thiết cho nhiều ứng dụng NLP khác như dịch máy, trả lời câu hỏi và nhận dạng thực thể.

Một khía cạnh đặc biệt thú vị của tiếng Nhật là sự tồn tại của các cấp độ lịch sự (keigo) khác nhau, được sử dụng để thể hiện sự tôn trọng và thân mật trong giao tiếp. Việc đánh giá mức độ lịch sự của một câu là một thách thức phức tạp, đòi hỏi phải hiểu ngữ cảnh, quan hệ xã hội của người nói và người nghe, cũng như sử dụng ngôn ngữ biểu cảm phù hợp.

Nghiên cứu gần đây đã thử nghiệm các mô hình máy học sâu, như mạng nơ-ron hồi quy và mô hình BERT, để giải quyết những thách thức này. Các mô hình này có thể học các đặc trưng ngữ nghĩa và ngữ cảnh phức tạp từ dữ liệu đào tạo lớn. Tuy nhiên, việc đánh giá lịch sự vẫn là một lĩnh vực nghiên cứu mở, đòi hỏi sự kết hợp của tri thức ngôn ngữ học, nhận thức xã hội và máy học.

Bằng cách phát triển thuật toán phân tích và đánh giá các câu tiếng Nhật một cách chính xác, từ đó xác định độ lịch sự của câu. Chúng em dùng bộ dữ liệu từ BCCWJ("The Balanced Corpus of Contemporary Written Japanese") và bộ phân tích cú pháp câu từ Ginza để lấy ra các loại từ đặc trưng. Điều này sẽ giúp cải thiện trải nghiệm người dùng, đồng thời thúc đẩy sự tôn trọng và lịch sự trong giao tiếp số.

2. Giới thiệu

Phân tích cú pháp câu là một phần không thể thiếu trong NLP, nhưng để đánh giá mức độ lịch sự của một câu là một nhiệm vụ thách thức đồng thời cũng rất cần thiết cho nhiều ứng dụng như bot chat trả lời tự động, phân tích tình cảm, hệ thống đối thoại,..

Trong giao tiếp tiếng Nhật, việc sử dụng thể lịch sự đóng vai trò quan trọng trong việc thể hiện sự tôn trọng và duy trì các mối quan hệ khác. Khái niệm lịch sự bao gồm nhiều cấp độ khác nhau, từ ngôn ngữ lịch sự đối với người cấp trên, đồng nghiệp đến ngôn ngữ khiêm tốn khi nói về bản thân. Sử dụng đúng mức độ lịch sự phù hợp trong ngữ cảnh là một yêu cầu quan trọng trong văn hóa giao tiếp của người Nhật.

Ở nghiên cứu này sẽ tập trung vào thuật toán dựa trên các quy tắc và hệ thống từ vựng, ngữ pháp tiếng Nhật để giải quyết việc đánh giá độ lịch sự. Nghiên cứu này mong muốn đóng góp một giải pháp hiệu quả và mở rộng, hỗ trợ nâng cao khả năng tiếng Nhật cho người dùng khi muốn biết một câu nói có độ lịch sự như nào.

3. Các nghiên cứu trước đây

- **Về BCCWJ (The Balanced Corpus of Contemporary Written Japanese)** [1]

Là một bộ ngữ liệu được tạo ra với mục đích nỗ lực nắm bắt tính đa dạng của tiếng Nhật viết hiện đại, bao gồm các mẫu văn bản tiếng Nhật hiện đại đa dạng để tạo ra một bộ ngữ liệu cân bằng độc đáo nhất có thể. Dữ liệu gồm 104,3 triệu từ, bao gồm các thể loại như sách và tạp chí chung, báo chí, báo cáo kinh doanh, blog, diễn đàn trực tuyến, sách giáo khoa và tài liệu pháp lý, v.v. Các mẫu ngẫu nhiên của từng thể loại đã được lấy.

Được phân loại theo các lý thuyết về phân loại văn bản, với các tiêu chí như chủ đề, phong cách viết, loại hình xuất bản, v.v. Điều này giúp bộ ngữ liệu có thể được sử dụng cho nhiều mục đích nghiên cứu khác nhau. Các mẫu được chọn ngẫu nhiên từ các nguồn khác nhau để đảm bảo tính đại diện và tránh thiên lệch. Phạm vi của bộ ngữ liệu là ~ 100 triệu từ (các mục từ vựng), không bao gồm khoảng trắng và ký hiệu. Được gán

[1] https://huggingface.co/datasets/allenai/c4

nhãn hình thái học và cú pháp dựa trên các khung lý thuyết và tập quy tắc phân tích ngôn ngữ tiếng Nhật. Điều này làm cho bộ ngữ liệu trở thành một nguồn tài nguyên giàu thông tin cho các ứng dụng xử lý ngôn ngữ tự nhiên.

- **Về GiNZA((Japanese Morphological Analysis System) là một công cụ xử lý ngôn ngữ tự nhiên nguồn mở cho tiếng Nhật, được phát triển bởi NHK (Tổng công ty Phát thanh Truyền hình Nhật Bản))[2]**

GiNZA sử dụng mô hình CRF (Conditional Random Field) và RNN (Recurrent Neural Network) để phân tích hình thái học (Morphological analysic) của câu tiếng Nhật thành các từ và gán nhãn từ loại

Phân tích cú pháp trong GiNZA dựa trên phương pháp Parser dựa trên chuyển tiếp (Transition-based Parser) sử dụng Stack LSTM. Nó phân tích cấu trúc phụ thuộc của câu thành các quan hệ phụ thuộc phân tán.

Một câu ví dụ cho việc GiNZA phân tích câu. Ở *Bảng 1* dưới đây là kết quả khi chạy phân tích cú pháp câu bằng GiNZA, cho ra thể từ điển của token đó, loại từ cũng như cách đọc. Đồng thời, công cụ cũng cho ra mối quan hệ của token để vẽ ParserTree.

Vì trong tiếng Nhật, trong một câu, các từ sẽ không được tách ra 1 khoảng trắng để người học có thể phân biệt từ vựng, ngữ nghĩa trong câu. Đồng thời công cụ cũng sẽ cho ra từ thể từ điển của từ đó. Giúp người học dễ dàng tiếp cận hơn với tiếng Nhật.

[2] https://github.com/megagonlabs/ginza

ぜひまた利用したいと思います。

(Nhất định tôi sẽ sử dụng nó một lần nữa)

Index	Chuỗi gốc	Gốc token (thể từ điển)	Chuẩn hóa	Cách đọc	Loại từ	Biến cách	Loại từ(tiếng Nhật)	Mối quan hệ phân tích cú pháp phụ thuộc của token	Index của token đầu trong mối quan hệ phân tích cú pháp phụ thuộc
0	ぜひ	ぜひ	是非	['ゼヒ']	ADV	[]	副詞	advmod	2
1	また	また	又	['マタ']	ADV	[]	副詞	advmod	2
2	利用	利用	利用	['リヨウ']	VERB	[]	名詞-普通名詞-サ変可能	ccomp	6
3	し	する	為る	['シ']	AUX	['サ行変格;連用形-一般']	動詞-非自立可能	aux	2
4	たい	たい	たい	['タイ']	AUX	['助動詞-タイ;終止形-一般']	助動詞	aux	2
5	と	と	と	['ト']	ADP	[]	助詞-格助詞	case	2
6	思い	思う	思う	['オモイ']	VERB	['五段-ワア行;連用形-一般']	動詞-一般	ROOT	6
7	ます	ます	ます	['マス']	AUX	['助動詞-マス;終止形-一般']	助動詞	aux	6
8	。	。	。	['。']	PUNCT	[]	補助記号-句点	punct	6

Bảng 1

4. Thuật toán đánh giá độ lịch sự

Dựa theo ngữ pháp tiếng Nhật và các từ lịch sự cũng như không lịch sự để viết thuật toán.

Thuật toán dùng sau khi câu tiếng nhật được xử lí token qua GiNZA, dò từng token so sánh với bộ từ và câu để đánh giá độ lịch sự. Các từ tiền tố mang tính lịch sự: "お", "ご". Các từ mang tính lịch sự: "ドウゾ","ドウモ","デス","マス". Đồng thời cũng có các từ mang tính không lịch sự.

Ở thuật toán này, sẽ duyệt qua từng token để xem trong câu có trùng từ nào nếu lịch sự thì tăng 1, ngược lại thì trừ 1. Tất nhiên trong tiếng Nhật sẽ có những từ như "いただきます","おじゃまします","お邪魔します","はいけんします","はいけんします", thì sẽ tính 5 (độ lịch sự cao nhất trong tiếng nhật)

Lấy ví dụ ở câu trên đã được parser, thuật toán này sẽ bắt được token “ます” (đây là 1 dạng ngữ pháp lịch sự trong tiếng nhật) khi đó sẽ tính mức độ lịch sự tăng thêm 1. Trong này sẽ có 11 mức độ thể hiện độ lịch sự từ vô cùng lịch sự đến rất thô lỗ (tương đương mức 5 đến -5)

5. Kết quả

a. Mức độ 5

“今日はトピックを発表させていただきます。”

(Hôm nay tôi xin phép thuyết trình về topic)

b. Mức độ 4

“みなさま、よるしくないお方々のようでございます。”

(Có vẻ mọi người có tâm trạng không tốt)

c. Mức độ 3

“あなたのようなタイプのお方が好きます。”

(Tôi thích tuýp người như bạn)

d. Mức độ 2

“木村さんのような人が口を利くべきではありません。”

(Người như anh Kimura thì không nên nói chuyện)

e. Mức độ 1

“その意味がよく分かりません。”

(Tôi không hiểu nó có ý nghĩa gì)

f. Mức độ 0

“年寄りなのに生意気だ。”

(Dù già nhưng trông vẫn rất khỏe mạnh)

g. Mức độ -1

“おまえはうそつき！”

(Mày là thằng nói dối!)

h. Mức độ -2

“おまえがやったのか、バカヤロー。”

(Là mày làm phải không, thằng đần)

i. Mức độ -5

“馬鹿ども”

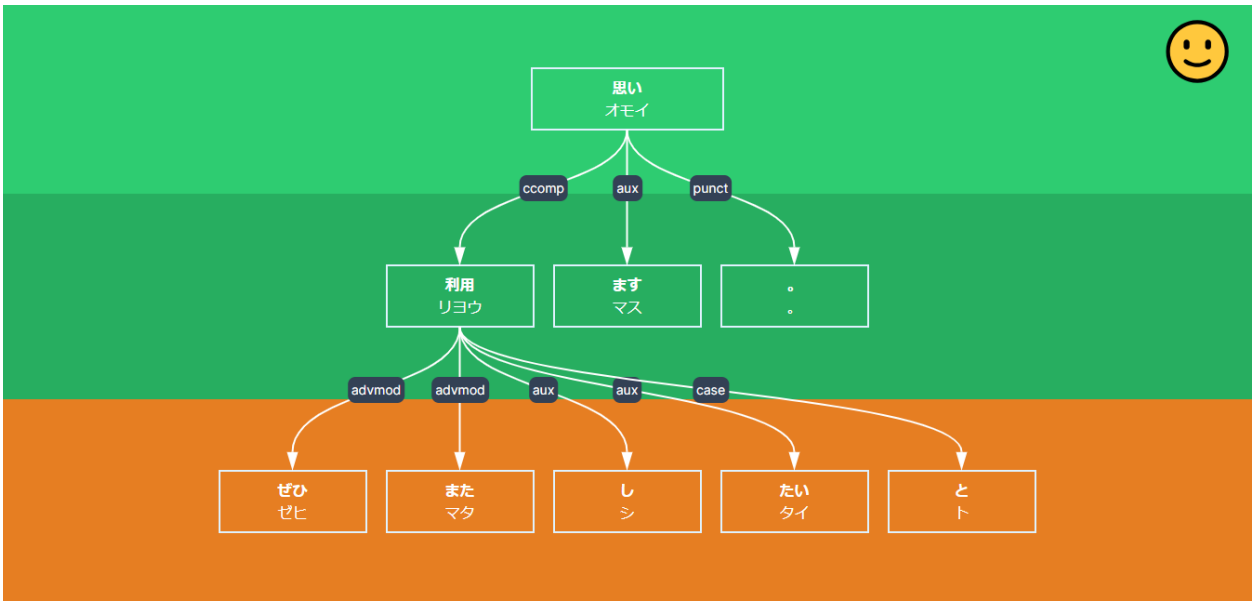
(Thằng ngu)

Việc này sẽ giúp người mới bắt đầu học tiếng Nhật cũng như là những người đã đi làm có sử dụng tiếng Nhật muốn sử dụng các câu lịch sự trong vấn đề giao tiếp, ứng xử trong môi trường học tập, làm việc để có thể phát triển bản thân cũng như công việc.

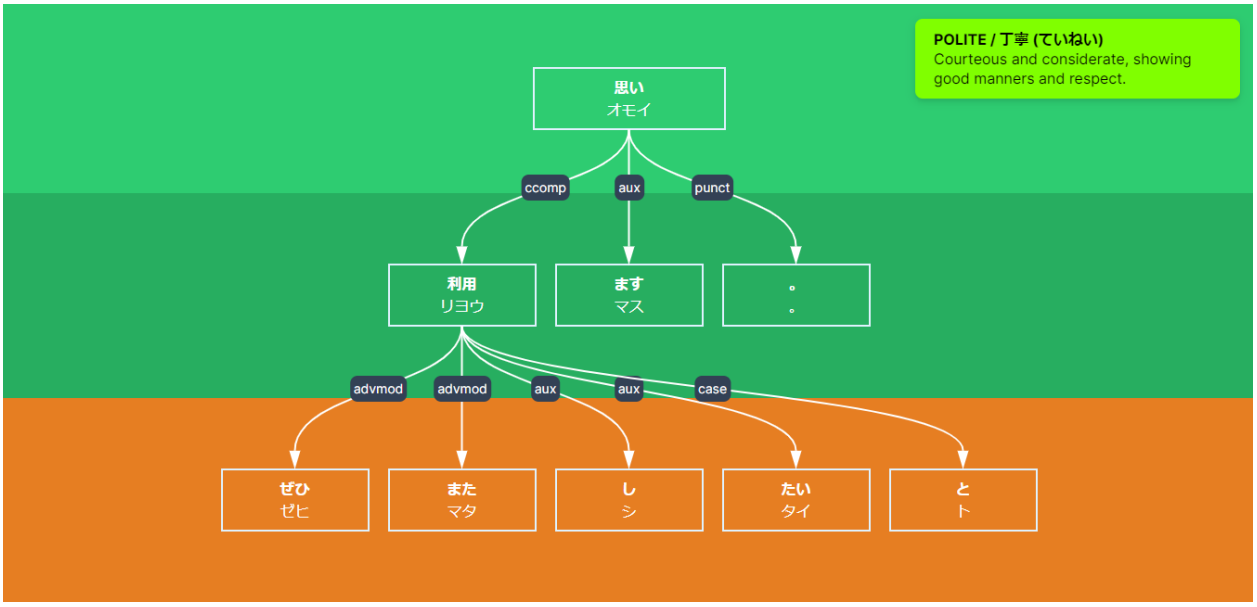
6. Ứng dụng

Dựa vào công cụ GiNZA và thuật toán, chúng em sẽ làm web để mô phỏng ParseTree

Dưới đây là web mô phỏng [3], ở bên góc phải có hiện icon khi nhấn vào sẽ hiện mức độ lịch sự. (lấy ví dụ ở *bảng 1*)



Hình 1



Hình 2

[3] https://wawakari.web.app/

7. Tài liệu tham khảo

1. https://study.gaijinpot.com/lesson/online-lessons/politeness-in-japanese/
2. https://huggingface.co/datasets/allenai/c4
3. https://github.com/megagonlabs/ginza
4. https://github.com/UniversalDependencies/UD_Japanese-BCCWJ
5. https://gengo.com/language-and-culture/polite-japanese-phrases-must-know/