**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**

--------------



**KHOÁ LUẬN TỐT NGHIỆP**

**XÂY DỰNG TRANG WEB HỖ TRỢ HỌC ĐỌC HIỂU**

**TIẾNG NHẬT ỨNG DỤNG CÁC KĨ THUẬT XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

**Giảng viên hướng dẫn: Ts. NGUYỄN LƯU THÙY NGÂN**

**Sinh viên thực hiện: TRẦN CÔNG THỨC**

**Lớp: KHTN**

**Khóa: 2015**

***Tp. Hồ Chí Minh, ngày 30 tháng 4 năm 2019***

**LỜI CẢM ƠN**

Trong suốt thời gian làm đề tài, em đã nhận được nhiều sự quan tâm giúp đỡ của giảng viên hướng dẫn. Với lòng biết ơn sâu sắc, em xin gởi lời cảm ơn chân thành đến cô Nguyễn Lưu Thùy Ngân – người đã tận tình dạy bảo và hướng dẫn cho em.

Vì thời gian và năng lực hạn chế nên trong quá trình làm báo cáo khó tránh khỏi sai sót về nội dung, lỗi trình bày mong được cô bỏ qua. Đồng thời em cũng mong nhận được ý kiến chia sẻ, đóng góp của các cô. Cuối cùng em xin kính chúc cô dồi dào sức khỏe và thành công trong sự nghiệp cũng như cuộc sống.

Trân trọng kính chào

**NHẬN XÉT**

**(Của giảng viên hướng dẫn)**

**MỤC LỤC**

[**I.** **MỞ ĐẦU** 4](#_Toc535530366)

[**II.** **TỔNG QUAN** 5](#_Toc535530367)

[**III.** **CÔNG CỤ KUROMOJI VÀ THUẬT TOÁN VITERBI** 6](#_Toc535530368)

[**1.** **Thư viện Kuromoji** 6](#_Toc535530369)

[**2.** **So sánh với các công cụ khác** 6](#_Toc535530370)

[**3.** **Thuật toán Viterbi** 6](#_Toc535530371)

[**IV.** **GIỚI THIỆU ỨNG DỤNG** 8](#_Toc535530372)

[**V.** **KẾT LUẬN** 9](#_Toc535530373)

[**VI.** **HƯỚNG PHÁT TRIỂN** 10](#_Toc535530374)

[**VII.** **TÀI LIỆU THAM KHẢO** 11](#_Toc535530375)

1. *Thêm mục: so sánh với website khác về tính năng*
2. *So sánh với chasen, về thuật toán, tốc độ,...*
3. *Giải thích về 2 bảng chữ cái, về đối tượng hướng đến //*

*C1: Tổng quan về ứng dụng*

1. *Tiếng Nhật và những khó khăn thường gặp phải*
2. *Đối tượng sử dụng*

*C2: Giới thiệu về ứng dụng*

1. *Giao diện*
2. *Hướng sử dụng*
3. *Lợi ích trong việc sử dụng*

*C3: Công cụ và thuật toán được sử dụng*

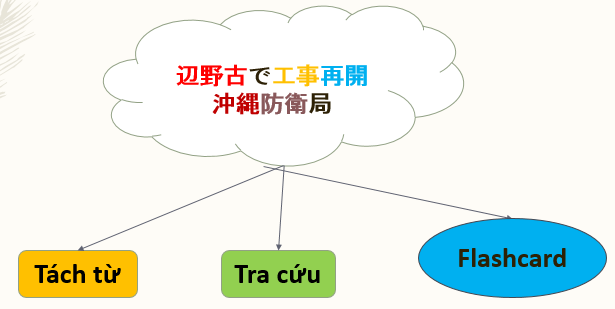
1. *Thư viện*
2. *So sánh*
3. *Thuật toán*
4. *Giới thiệu ứng dụng*
5. *Tổng quan về tiếng Nhật*
6. *Đối tượng sử dụng*
7. Ngoài ra, việc các từ không được phân chia theo khoảng trắng sẽ khiến người đọc khó xác định rõ ràng ranh giới giữa các từ, nhất là với những cụm từ ghép Kanji có độ dài lớn
8. với chức năng tách các từ để người đọc dễ dàng phân biệt giữa các từ đó, hỗ trợ tìm hiểu và giúp họ học các từ mới bằng chức năng tạo flashcard
9. **MỞ ĐẦU**

Trong bối cảnh hội nhập của Việt Nam hiện nay, việc học ngoại ngữ là điều thực sự rất cần thiết, giúp mở ra thêm nhiều những cơ hội học tập, làm việc với các nhiều nước trên thế giới. Với việc đông đảo các doanh nghiệp Nhật Bản đã và đang đầu tư rất nhiều vào Việt Nam ở nhiều lĩnh vực khác nhau, thì xu hướng học tiếng Nhật trong giới trẻ hiện nay lại ngày càng sôi động.

Vì tiếng Nhật là một ngôn ngữ khó nên để có thể thành thạo thì việc học cần có các phương pháp cụ thể và hiệu quả. Một trong số đó là học thông qua các bài báo. Với việc đọc báo, ta sẽ luyện được kĩ năng đọc hiểu, đồng thời học thêm được các từ vừng mới cũng như ôn luyện lại những từ vựng cũ đã học. Ứng dụng web hỗ trợ đọc báo tiếng Nhật được tạo ra nhằm hỗ trợ người đọc những kĩ năng như vậy, giúp cho việc đọc báo được dễ dàng hơn và việc học tiếng Nhật cũng trở nên thú vị hơn.

1. **TỔNG QUAN**

Đối với những người đang trong quá trình học tiếng Nhật, việc đọc hiểu văn bản là một việc khó khăn khi bảng chữ cái tiếng Nhật khác hoàn toàn với bảng chữ cái Latinh vốn đã quen thuộc với người Việt. Ngoài ra, cấu trúc câu không phân chia các từ theo khoảng trắng sẽ khiến người đọc khó xác định rõ ràng ranh giới giữa các từ, nhất là với những cụm từ ghép Kanji có độ dài lớn. Ứng dụng web hỗ trợ đọc báo tiếng Nhật với chức năng tách các từ để người đọc dễ dàng phân biệt giữa các từ đó, hỗ trợ tìm hiểu và giúp họ học các từ mới bằng chức năng tạo flashcard. Trên internet đã có một số trang đọc báo có hỗ trợ tách và đánh dấu từ mới, nhưng tất cả đều cố định trên một trang báo từ trước, không tách toàn bộ văn bản và cũng không cho phép người dùng đánh dấu và tìm hiểu một từ mới. Với tính năng tạo flashcard, ứng dụng có tính thực tiễn cao hơn khi giúp đỡ người đọc đánh dấu để học thêm được từ mới, hướng tới mục tiêu hoàn thiện kĩ năng đọc hiểu văn bản tiếng Nhật thông qua việc đọc báo



1. **CÔNG CỤ KUROMOJI VÀ THUẬT TOÁN VITERBI**
2. **Thư viện Kuromoji**

Kuromoji là một công cụ phân tích hình thái tiếng Nhật được viết bằng ngôn ngữ Java. Nó hỗ trợ các tính năng như:

* Phân chia đoạn văn bản tiếng Nhật thành các từ riêng lẻ
* Phân loại từ theo danh mục (danh từ, tính từ, động từ ,...)
* Truy xuất nguyên mẫu của động từ và tính từ
* Truy xuất cách đọc của từ

1. **So sánh với các công cụ khác**

So với một số công cụ khác, Kuromoji có đầy đủ các tính năng hỗ người dùng phân tích hình thái từ trong các đoạn văn bản tiếng Nhật

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Kuromoji | Mecab | KyTea | Juman |
| Phân đoạn từ | x | X | x | x |
| Phân loại từ | x | X | x | x |
| Truy xuất nguyên mẫu | x | X |  |  |
| Truy xuất cách đọc | x | X | x |  |

So với Mecab là một công cụ phân tích hình thái tiếng Nhật khác cũng khá phổ biến, Kuromoji có tốc độ tương đương do có cùng sử dụng thuật toán Viterbi để đánh giá khả năng phân đoạn từ

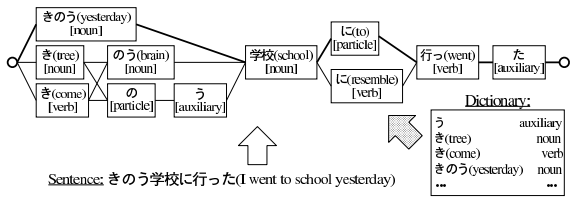
1. **Thuật toán Viterbi**

Với mỗi đoạn văn bản n kí tự, kiểu chuỗi C = c1 c2 c3... cn, có nhiều hơn 1 cách phân đoạn thành m thành phần W= w1 w2 w3... wm. Mục tiêu của thuật toán là tìm cách phân đoạn mà có xác suất xảy ra lớn nhất:



Với  là xác suất từ  xảy ra nếu trước đó là từ. Xác suất này được tính như sau: (các phép đếm được liên kết với từ điển MeCab-IPADIC)

Với các trong mỗi cách phân đoạn W, ta dựng được mô hình như bên dưới, với mỗi node là một từ



Thuật giải

best\_score[0] = 0

for each node in the graph (ascending order)

best\_score[node] = ∞

for each incoming edge of node

score = best\_score[edge.prev\_node] + edge.score

if score < best\_score[node]

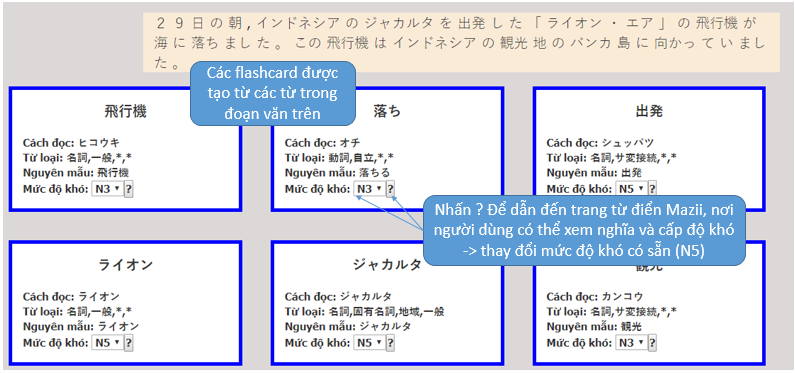
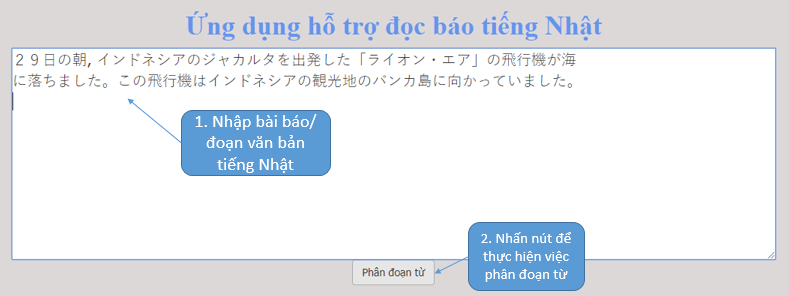
best\_score[node] = score

Trong đó: best\_score[i]: xác suất lớn nhất để phân đoạn từ, tính đến node i

edge.score: xác suất Pr giữa 2 node

Thực hiện thuật toán cho đến khi đã xét hết các node, kết quả cuối cùng là cách phân đoạn có xác suất lớn nhất

1. **GIỚI THIỆU ỨNG DỤNG**



1. **KẾT LUẬN**

Bước đầu, ứng dụng đã giúp người dùng phân đoạn các đoạn văn thành từ riêng lẻ, đồng thời giúp người dùng xem một số thuộc tính của từ (cách đọc, từ loại, nguyên mẫu) một cách nhanh chóng, hỗ trợ liên kết với từ điển khác để xem được nghĩa và mức độ khó để người dùng có thể phân loại

Do vẫn trong giai đoạn hoàn thiện, ứng dụng còn một số những thiếu sót, cần được bổ sung và hoàn thiện thêm, tiến đến mở rộng các chức năng để phục vụ người dùng một cách tốt hơn

1. **HƯỚNG PHÁT TRIỂN**

Để thuận tiện hơn cho người dùng, ứng dụng cần tích hợp từ điển để có thể tra từ một cách nhanh chóng mà không cần phải liên kết với trang từ điển khác. Đối với việc tạo flashcard, cần có những chức năng khác như xóa, download flashcard xuống máy tính, người dùng đăng nhập để có thể lưu flashcard ngay trên ứng dụng

1. **TÀI LIỆU THAM KHẢO**

Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information - Tetsuji Nakagawa

A two-stage statistical word segmentation system for Chinese - Guohong Fu & K.K. Luke

http://www.atilika.org/