**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**

**TRẦN CÔNG THỨC**

**KHÓA 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**

**Building a web-based application for improving Japanese reading comprehension skill using natural language processing techniques**

**CỬ NHÂN NGÀNH KHOA HỌC MÁY TÍNH**

**TP. HỒ CHÍ MINH, 2019**

**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**

**TRẦN CÔNG THỨC – 15520864**

**KHÓA 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**

**Building a web-based application for improving Japanese reading comprehension skill using natural language processing techniques**

**CỬ NHÂN NGÀNH KHOA HỌC MÁY TÍNH**

**GIẢNG VIÊN HƯỚNG DẪN**

**Tiến sĩ NGUYỄN LƯU THÙY NGÂN**

**TP. HỒ CHÍ MINH, 2019**

DANH SÁCH HỘI ĐỒNG BẢO VỆ KHÓA LUẬN

Hội đồng chấm khóa luận tốt nghiệp, thành lập theo Quyết định số …………………… ngày ………………….. của Hiệu trưởng Trường Đại học Công nghệ Thông tin.

* 1. …………………………………………. – Chủ tịch.
  2. …………………………………………. – Thư ký.
  3. …………………………………………. – Ủy viên.
  4. …………………………………………. – Ủy viên.

**MỤC LỤC**

[TÓM TẮT KHÓA LUẬN 1](#_Toc7512073)

[MỞ ĐẦU 2](#_Toc7512074)

[Chương 1. TỔNG QUAN 3](#_Toc7512075)

[1.1. Giới thiệu về tiếng Nhật 3](#_Toc7512076)

[1.2. Các ứng dụng có sẵn và khuyết điểm 6](#_Toc7512077)

[1.2.1. NHK NEW WEBS EASY 6](#_Toc7512078)

[1.2.2. Tangoristo 8](#_Toc7512079)

[1.3. Ưu điểm hướng đến của ứng dụng 9](#_Toc7512080)

[Chương 2. THƯ VIỆN KUROMOJI VÀ THUẬT TOÁN VITERBI 10](#_Toc7512081)

[2.1. Thư viện Kuromoji 10](#_Toc7512082)

[2.2. So sánh với các công cụ khác 10](#_Toc7512083)

[2.2.1. Chức năng hỗ trợ 10](#_Toc7512084)

[2.2.2. Tốc độ thực thi 10](#_Toc7512085)

[2.3. Thuật toán Viterbi 10](#_Toc7512086)

[Chương 3. GIỚI THIỆU VỀ ỨNG DỤNG 12](#_Toc7512087)

[3.1. Các công nghệ được sử dụng 12](#_Toc7512088)

[3.2. Sơ đồ ứng dụng 12](#_Toc7512089)

[3.3. Giao diện 12](#_Toc7512090)

[3.4. Đánh giá của người dùng 13](#_Toc7512091)

[Chương 4: KẾT LUẬN 14](#_Toc7512092)

[Chương 5: HƯỚNG PHÁT TRIỂN 15](#_Toc7512093)

[TÀI LIỆU THAM KHẢO 16](#_Toc7512094)

DANH MỤC HÌNH VẼ

**Chương 1:**

[Hình 1. 1: Bảng chữ Hiragana 3](#_Toc7503037)

[Hình 1. 2: Bảng chữ Katakana 4](#_Toc7503038)

[Hình 1. 3: Một số chữ Kanji 5](#_Toc7503039)

[Hình 1. 4: Trang NEW WEBS EASY 6](#_Toc7503040)

[Hình 1. 5: Ứng dụng Tangoristo 8](#_Toc7503041)

**Chương 2:**

[Hình 2. 1 : Sơ đồ minh họa thuật toán 11](#_Toc7506201)

**Chương 3:**

[Hình 3. 1: Giao diện ứng dụng – nhập text 12](#_Toc7502777)

[Hình 3. 2: Giao diện ứng dụng – nhập URL 13](#_Toc7502778)

[Hình 3. 3: Giao diện ứng dụng – các bước tìm hiểu 14](#_Toc7502779)

DANH MỤC BẢNG

[Bảng 2. 1: Các tính năng hỗ trợ của một số thư viện 10](#_Toc7506388)

TÓM TẮT KHÓA LUẬN

Đối với những người đang trong quá trình học tiếng Nhật, việc đọc hiểu văn bản là một việc khó khăn khi tiếng Nhật là một ngôn ngữ tượng hình với một lượng lớn các từ ngữ khác hoàn toàn với bảng chữ cái Latin vốn đã quen thuộc với người Việt. Khóa luận hướng tới việc nghiên cứu và xây dựng một ứng dụng web hỗ trợ đọc hiểu tiếng Nhật dựa trên những công cụ có sẵn, với các chức năng mang tính thực tiễn cao, hướng tới giúp người dùng dễ dàng hơn trong việc học và hoàn thiện kĩ năng tiếng Nhật của mình một cách hiệu quả.

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 trong nhiều lĩnh vực khác nhau, đồng thời cộng với việc các cơ hội được sang Nhật Bản học tập và làm việc ngày càng được mở rộng thì việc học tiếng Nhật trở thành một nhu cầu quan trọng đối với đông đảo người Việt Nam hiện nay

Tiếng Nhật vốn 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 hiểu được dễ dàng hơn và việc học tiếng Nhật cũng trở nên thú vị hơn. Ứng dụng hướng đến những người có những hiểu biết cơ bản về tiếng Nhật, về cấu trúc câu và 2 bộ chữ cơ bản là Hiragana và Katakana, đang gặp khó khăn trong quá trình tiếp cận những đoạn văn bản phức tạp hơn với số lượng Kanji nhiều và khó. Việc sử dụng ứng dụng sẽ là một công cụ trợ giúp hữu hiệu trong việc giúp người dùng giải quyết những khó khăn đó và có được hiểu quả cao hơn trong việc đọc hiểu cũng như là học tiếng Nhật

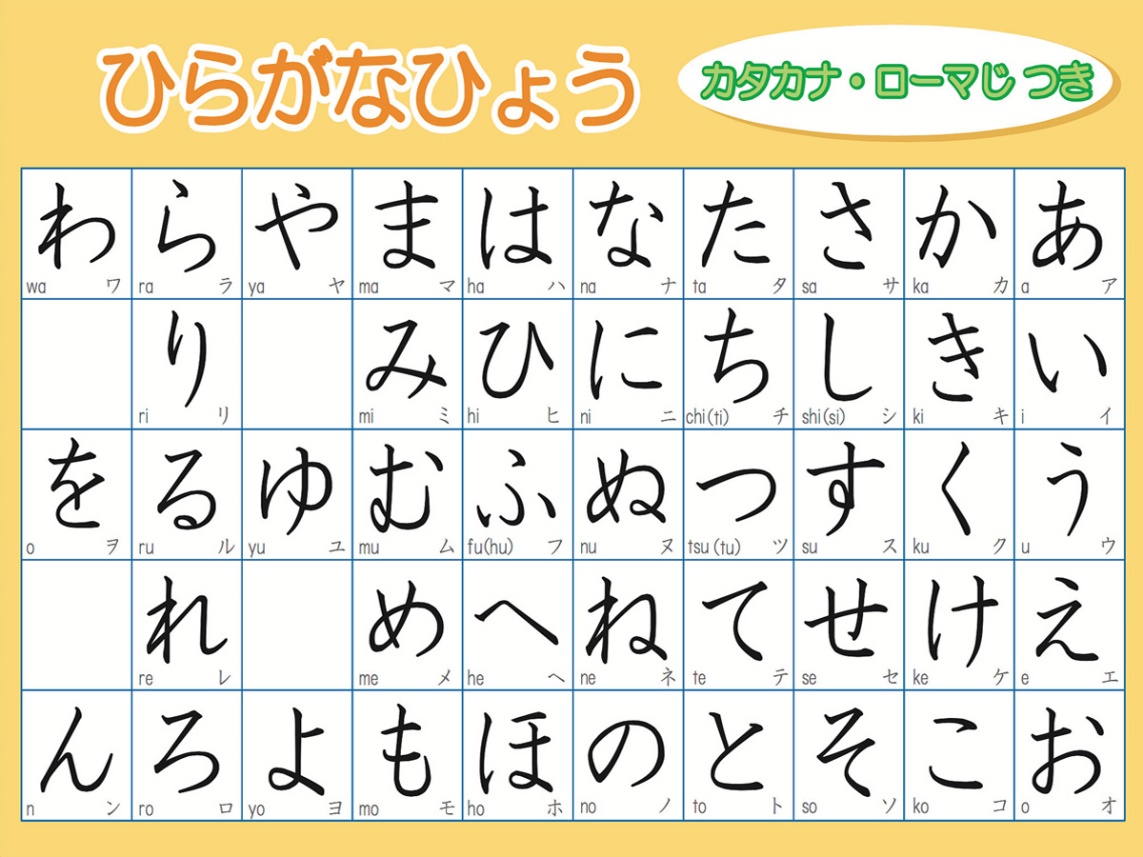
# TỔNG QUAN

## Giới thiệu về tiếng Nhật

**Tiếng Nhật Bản**, **tiếng Nhật** hay **Nhật ngữ** là một [ngôn ngữ Đông Á](https://vi.wikipedia.org/wiki/C%C3%A1c_ng%C3%B4n_ng%E1%BB%AF_%C4%90%C3%B4ng_%C3%81) được hơn 130 triệu người sử dụng ở [Nhật Bản](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADt_B%E1%BA%A3n) và những cộng đồng dân di cư [Nhật Bản](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADt_B%E1%BA%A3n) khắp thế giới.

Về tổng thể, tiếng Nhật được viết trong sự phối hợp giữa 3 kiểu chữ: đơn âm mềm Hiragana, đơn âm cứng Katakana và Kanji (Hán tự kiểu Nhật).

**Hiragana （ひらがな）[3]** còn được gọi là chữ mềm (nét chữ mềm mại), là dạng ký tự truyền thống của người Nhật, được phát triển từ chữ Hán để diễn đạt cách phát âm của người Nhật. Bộ Hiragana bao gồm 46 chữ cái, được xem là chữ thuần Nhật, được dùng để viết từ chỉ có trong tiếng Nhật, đuôi của động từ và từ khó đọc, khó nhớ bằng Kanji.



Hình 1. 1: Bảng chữ Hiragana

**Katakana （カタカナ）[3]** còn được gọi là chữ cứng (các chữ có nét cong, thẳng, gấp khúc), được dùng để phiên âm những từ vay mượn từ tiếng nước ngoài. Ví dụ, camera sẽ được viết là カメラ (Kamera), taxi là タクシー (takushi). Cũng với tính chất đó, chữ Katakana còn được dùng để viết tên quốc gia, địa danh hay tên người nước ngoài.



Hình 1. 2: Bảng chữ Katakana

**Kanji （漢字）[3]**là loại chữ tượng hình mượn từ chữ Hán, được cho là du nhập vào Nhật từ thế kỷ thứ 5. Những Hán tự này ban đầu cũng được đọc bằng tiếng Hán, về sau được đọc theo quy tắc ngữ pháp tiếng Nhật.

Nếu Hiragana và Katakana có cùng điểm chung đó là mỗi chữ biểu thị một âm tiết, tức là chỉ có một cách đọc, thì Kanji lại có nhiều cách đọc tùy thuộc vào hoàn cảnh. Trong tiếng Nhật hiện đại, Kanji được dùng để viết một số phần trong câu, như danh từ, tính từ, động từ (đa số nhưng không phải tất cả các từ loại này đều được viết bằng Kanji). Kanji thường được sử dụng trong văn bản, tài liệu (những trường hợp chính thống) xen kẽ với những bộ chữ khác.



Hình 1. 3: Một số chữ Kanji

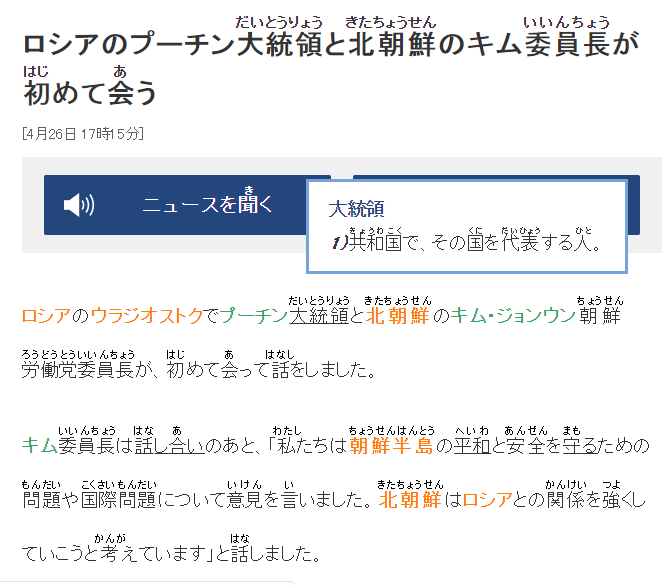
Thoạt nhìn, tiếng Nhật có cách viết phức tạp khi kết hợp cả 3 bộ chữ trong cùng một câu. Nhưng cũng nhờ vậy mà người đọc sẽ dễ phân biệt các thành phần và hiểu ý nghĩa của câu nhanh hơn.

Ngoài 3 bộ chữ nêu trên, tiếng Nhật còn được biểu thị bằng một dạng thứ 4 là **Romaji,** đây là cách dùng chữ Latinh để phiên âm cách đọc tiếng Nhật, thường dùng cho những người nước ngoài mới học tiếng và chưa thể đọc được các ký tự trong 3 bộ chữ Hiragana, Katakana và Kanji.

Với 2 bộ chữ Hiragana và Katakana, việc kết cấu của chữ đơn giản cộng với việc số lượng ít (46 chữ) thì việc đọc và học 2 bộ chữ này đều không khó và chỉ mất một khoảng thời gian ngắn. Tuy nhiên, đối với chữ Kanji, số lượng chữ lớn và kết cấu chữ là một trở ngại không hề nhỏ đối với người mới bắt đầu học tiếng Nhật. Tính đến hiện nay thì Kanji có tổng cộng khoảng 3000 chữ, trong đó có tầm 1500 – 1900 chữ thông dụng. Mỗi một chữ có thể là kết hợp của nhiều bộ chữ thành phần nhỏ hơn (gọi là bộ thủ) hoặc là của những nét chữ riêng biệt hợp thành. Đối với một từ ghép gồm 1 hay nhiều chữ Kanji hợp thành, việc không phân khoảng cách giữa các từ như chữ Latin sẽ gây khó khăn thêm cho người đọc trong việc phân biệt các từ, từ đó việc đọc văn bản sẽ trở thành một trở ngại không nhỏ trong quá trình học tiếng Nhật đối với mọi người, cần có nhiều thời gian, công sức cũng như một cách học và luyện tập hợp lý để cải thiện kĩ năng này.

## Các ứng dụng có sẵn và khuyết điểm

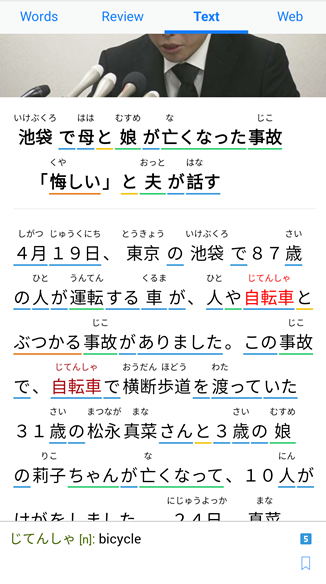
### NHK NEW WEBS EASY



Hình 1. 4: Trang NEW WEBS EASY

NEW WEBS EASY là một phiên bản của NHK NEWS WEB – một trang thông tin thời sự của đài phát thanh truyền hình NHK – Nhật Bản. Trang NEW WEBS EASY chuyên đăng tải những bản tin thời sự với cách diễn đạt và sử dụng từ ngữ tương đối đơn giản cho những người mới học tiếng Nhật có thể đọc hiểu tin tức một cách dễ dàng hơn. Các từ Kanji đều được ghi chú bằng chữ Hiragana ở trên đầu để cho người dùng dễ đọc và tra cứu, đồng thời khi rê chuột vào một số từ khó, những gợi ý về ý nghĩa của từ được đưa ra để cho người dùng dễ nắm bắt nghĩa của từ mà không cần tra từ điển. Tuy vậy, các bài báo, những gợi ý về từ đều có sẵn, khiến người dùng không linh động trong việc tra cứu những bài báo cũng như từ vựng mong muốn, đồng thời cấu trúc từ ngữ khá đơn giản, thích hợp hơn cho những người mới bắt đầu học đọc hiểu tiếng Nhật. Tính ứng dụng và tiện dụng của trang web không thực sự cao, mức độ kiến thức đạt được ở phạm vi nhỏ nên không thực sự phù hợp với nhiều đối tượng người dùng.

### Tangoristo



Hình 1. 5: Ứng dụng Tangoristo

Tangoristo là một ứng dụng trên điện thoại Android có hỗ trợ người dùng đọc hiểu tiếng Nhật thông qua các bài báo, với nguồn từ trang NHK. Ứng dụng có hỗ trợ chức năng phân tách các từ, gợi ý cách đọc, từ loại và nghĩa của từ. Tương tự như với trang web NEW WEBS EASY, ứng dụng Tangoristo vẫn chỉ hỗ trợ những bài báo cố định với số lượng gợi ý có sẵn và hạn chế, khiến người dùng không linh động trong việc đọc và tra cứu những từ vựng mong muốn

## Ưu điểm hướng đến của ứng dụng

Trang web hỗ trợ đọc báo tiếng Nhật có chức năng tách các đoạn văn bản tiếng Nhật thành các từ riêng biệt, giúp người dùng tìm hiểu những đặc điểm của từ đó (từ loại, nguyên mẫu, cách đọc), tạo flashcard để người dùng có thể lưu và học các từ mới trong quá trình sử dụng. Người dùng sẽ chủ động trong việc chọn các đoạn văn bản, bài báo cũng như các từ vựng cần tìm hiểu, giúp cho quá trình học đọc hiểu tiếng Nhật trở nên tiện lợi và hiệu quả hơn.

# THƯ VIỆN KUROMOJI VÀ THUẬT TOÁN VITERBI

## 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 [4]:

* 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

## So sánh với các công cụ khác

### Chức năng hỗ trợ

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 | Chasen | KyTea | Juman |
| Phân đoạn từ | X | X | X | X | X |
| Phân loại từ | X | X | X | X | X |
| Truy xuất nguyên mẫu | X | X | X |  |  |
| Truy xuất cách đọc | X | X | X | X |  |

Bảng 2. : Các tính năng hỗ trợ của một số thư viện

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

### Tốc độ thực thi

(Đang cập nhật)

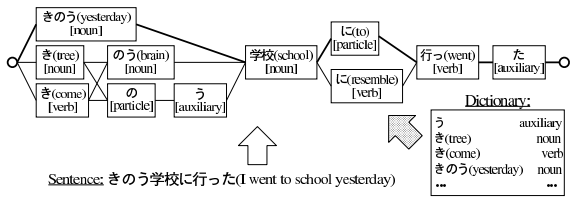
## Thuật toán Viterbi

[2] 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 [1], với mỗi node là một từ



Hình 2. : Sơ đồ minh họa thuật toán

**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

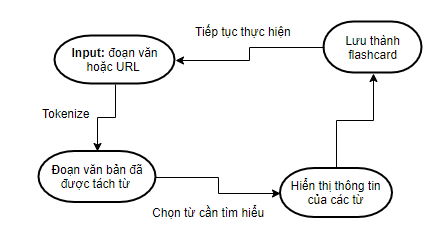
Thực hiện thuật toán cho đến khi đã xét hết các node, kết quả cuối cùng là cách phân đoạn có xác suất lớn nhất

# GIỚI THIỆU VỀ ỨNG DỤNG

## Các công nghệ được sử dụng

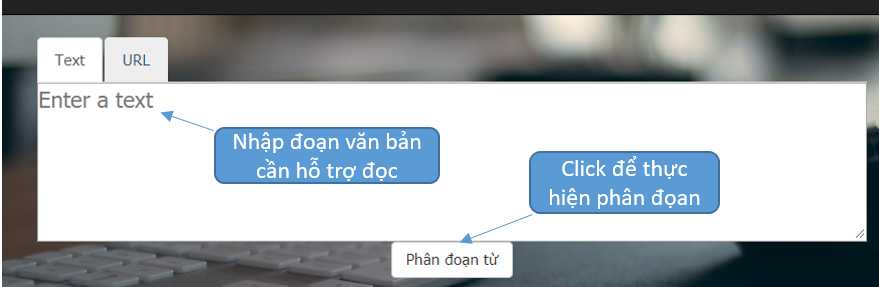
Ứng dụng Web hỗ trợ đọc hiểu tiếng Nhật được viết bằng ngôn ngữ Java, với sự kết hợp của các Framework Java như Spring MVC, Servlet ở phần BackEnd và HTML, CSS, Javascript, cùng với framework Jquery và Bootstrap ở phía FrontEnd

## Sơ đồ ứng dụng

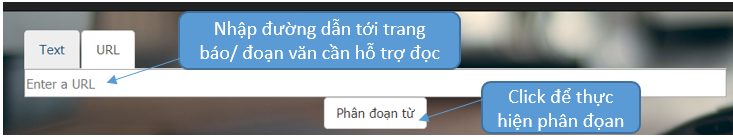


Hình 3. : Sơ đồ ứng dụng

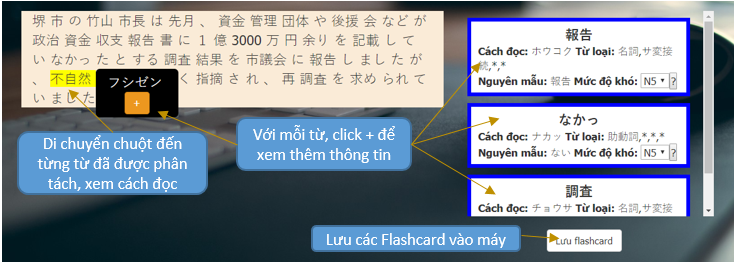
## Giao diện



Hình 3. 2: Giao diện ứng dụng – nhập text



Hình 3. 3: Giao diện ứng dụng – nhập URL



Hình 3. 4: Giao diện ứng dụng – các bước tìm hiểu

## Đánh giá của người dùng

(đang cập nhật)

Chương 4: 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. Thư viện hoạt động nhanh chóng và chính xác, đáp ứng được yêu cầu của ứng dụng

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

**Chương 5: HƯỚNG PHÁT TRIỂN**

Để thuận tiện hơn cho người dùng, ứng dụng cần tích hợp từ điển để có thể tra nghĩa từ một cách nhanh chóng mà không cần phải liên kết với trang từ điển khác, đồng thời biết về mức độ khó của từ để sắp xếp thời gian cũng như phương pháp học để học từ mới một cách hiệu quả. Đối với việc tạo flashcard, cần có những chức năng khác như xóa, sửa flashcard cũng như lưu trực tiếp các flashcard cũ trên tài khoản đăng nhập người dùng. Ngoài ra, vì tính ứng dụng cao, tiềm năng phát triển tốt nên ứng dụng cần được xây dựng với các công nghệ phù hợp để đáp ứng lượng người dùng lớn, hỗ trợ thêm tiếng Anh bên cạnh tiếng Việt để phục vụ người dùng từ những nước khác

TÀI LIỆU THAM KHẢO

[1] Tetsuji Nakagawa, “Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information”

[2] Guohong Fu & K.K. Luke, “A two-stage statistical word segmentation system for Chinese”

[3] Du học Hoa Sen, “Khái quát về tiếng Nhật”, Available: <https://duhochoasen.com/khai-quat-ve-tieng-nha%CC%A3t.html>

[4] Atilika Company, “Kuromoji”, Available: <https://www.atilika.com/en/kuromoji/>