TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN CÔNG NGHỆ TRI THỨC

NGUYỄN NGỌC GIA
HY - 1212166

SỬ DỤNG LATEX TRONG KHOÁ LUẬN TỐT NGHIỆP

KHOÁ LUẬN TỐT NGHIỆP CỬ NHÂN CNTT

GIÁO VIÊN HƯỚNG DẪN
ĐINH ĐIỀN

KHOÁ 2012-2016

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

Đây là đề tài có ý nghĩa thực tiễn và hữu ích khi ứng dụng trong thực tế. Sinh viên đã biết kết hợp điểm mạnh của từng mô hình (noisy channel và độ liên kết ngữ cảnh) để cho ra kết quả tốt nhất có thể được. Sinh viên có đề xuất sử dụng ngưỡng động và hàm max trong cách tính độ liên kết ngữ cảnh và đã chứng minh sự hiệu quả của đề xuất này dựa trên kết quả thực nghiệm.

Cách trình bày của khoá luận tốt nghiệp tuân thủ theo cách trình bày chuẩn, khá súc tích, ngắn gọn.

Trong quá trình làm, sinh viên có thể hiện sự cố gắng để hoàn thành khoá luận tốt nghiệp một cách tốt nhất trong khả năng của mình. Tuy đôi khi vẫn còn sắp xếp thời gian chưa hợp lí nhưng nhìn chung sinh viên đã hoàn thành tốt so với yêu cầu của một khoá luận tốt nghiệp.

Tp. HCM, ngày 21 tháng 07 năm 2015
Giáo viên hướng dẫn

Đinh Điền

NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

Bài toán kiểm lỗi chính tả là một trong những bài toán cơ bản và quen thuộc trong xử lý ngôn ngữ tự nhiên. Giải quyết bài toán này giúp cho người sử dụng máy tính tránh được những lỗi văn bản không đáng có do nhầm lẫn khi gõ trên bàn phím hay do thói quen phát âm chưa chuẩn,...

Trong xử lý ngôn ngữ tự nhiên tiếng Việt đã có những công trình giải quyết bài toán này trong những năm qua, mặc dù vẫn còn lẻ tẻ. Khoá luận tốt nghiệp này đóng góp thêm một lời giải cho bài toán này.

Rõ ràng rằng việc phát hiện ra lỗi có thể đơn giản trong trường hợp tiếng không xuất hiện trong từ điển (lỗi non-syllable) nhưng khi tiếng đó thực sự tồn tại thì việc quyết định đó phải là lỗi sai hay không phụ thuộc nhiều vào ngữ cảnh các tiếng xung quanh. Hơn nữa, bài toán này không chỉ dừng ở mức phát hiện lỗi mà phải đến mức đề xuất phương án sửa lỗi sai đó.

Khoá luận này đã chọn được hướng tiếp cận thích hợp khi sử dụng liên kết ngữ cảnh để xác định lỗi sai. Đồng thời lựa chọn ứng viên thích hợp từ tập ứng viên phát sinh được để đề xuất cách sửa lỗi.

Khoá luận này được phát triển trên cơ sở của nhóm tác giả Nguyễn Thị Xuân Hương và có thêm cải tiến nhờ vào sự quan sát, lý luận và thử nghiệm. Đóng góp chính của khoá luận là thay đổi hàm quyết định trong liên kết ngữ cảnh (dùng hàm max thay vì hàm trung bình nhân) và đưa ra ngưỡng động (thay vì ngưỡng tĩnh) trên cơ sở của lý thuyết Noisy Channel.

Tác giả khoá luận đã dành nhiều công sức thu thập và xây dựng ngữ liệu hợp lý (trên sự hỗ trợ của các cộng tác viên) để bước đầu có thể dùng để thử nghiệm và đánh giá. Kết quả thử nghiệm cho thấy các đề xuất và cải tiến trong khoá luận là đáng ghi nhận và có thể chấp nhận được.

Tác giả khoá luận đã trình bày báo cáo khoá luận gồm 6 chương và 2 phụ lục rõ ràng và hợp lý. Dĩ nhiên rằng, báo cáo hiện giờ tốt hơn nhiều so với báo cáo ban đầu sau khi gặp người phản biện. Điều đó cho thấy tác giả đã lắng nghe để hoàn chỉnh khoá luận này về mặt báo cáo và chương trình.

Điểm hạn chế của khoá luận nằm ở chỗ giả định lỗi sai xuất hiện chỉ tại một tiếng khi xem xét. Trên thực tế lỗi sai có thể xuất hiện từ hai tiếng liền kề trở lên và giữa các lỗi sai có tác động qua lại lẫn nhau.

Ngoài ra, do hướng tiếp cận dựa trên thống kê khiến cho những từ hiếm gặp hoặc thường sai do thói quen của đại đa số người không thể phát hiện lỗi được (chẳng hạn với cụm từ "vô hình trung").

Dù vậy, tác giả đã thể hiện sự cố gắng, nền tảng kiến thức và các kỹ năng của mình để hoàn thành khoá luận tốt nghiệp đáp ứng yêu cầu một khoá luận tốt nghiệp cử nhân Công nghệ thông tin.

Khóa luận đáp ứng yêu cầu của Khóa luận cử nhân CNTT.

Tp. HCM, ngày 21 tháng 07 năm 2015

Giáo viên phản biện

Nghiêm Quốc Minh

Lời cảm ơn

Tôi xin chân thành cảm ơn ...

Khoa Công Nghệ Thông Tin
Bộ môn Công nghệ tri thức

ĐỀ CƯƠNG CHI TIẾT

Tên đề tài: Sử dụng LaTeX trong Khoá luận tốt nghiệp
Giáo viên hướng dẫn: Đinh Điền
Thời gian thực hiện: 01/01/2000-01/01/2001
Sinh viên thực hiện: Nguyễn Ngọc Gia Hy - 1212166
Loại đề tài: Tìm hiểu công nghệ (có hoặc không ứng dụng minh hoạ), Xây dựng ứng dụng, ...

Nội dung đề tài: mô tả chi tiết nội dung đề tài, yêu cầu, phương pháp thực hiện, kết quả đạt được, ...
Kế hoạch thực hiện: mô tả chi tiết thời gian của các giai đoạn thực hiện và phân công công việc của từng thành viên trong nhóm

Xác nhận của GVHD

Ngày ... tháng ... năm 2015

Đinh Điền

Nguyễn Ngọc Gia Hy

Mục lục

Danh sách hình

Danh sách bảng

Tóm tắt

Tóm tắt khóa luận: trình bày tóm tắt vấn đề nghiên cứu, các hướng tiếp cận, cách giải quyết vấn đề và một số kết quả đạt được.

Chương 1

Giới thiệu chung

Ngôn ngữ để viết và trình bày luận văn là tiếng Việt hoặc tiếng Anh Anh văn. Trường hợp chọn ngôn ngữ tiếng Anh để viết và trình bày luận án, học viên cao học (HVCH) cần có văn bản đề nghị, được cán bộ hướng dẫn (CBHD) đồng ý và nộp cho phòng Đào tạo Sau đại học (phòng ĐT SĐH) vào thời điểm đăng ký đề tài luận văn để xin ý kiến phê duyệt của Thủ trưởng cơ sở đào tạo (CSĐT). Luận văn viết và trình bày bằng tiếng Anh phải có bản tóm tắt luận văn viết bằng tiếng Việt.

Tóm tắt luận văn: Tóm tắt luận văn phải in theo kích thước 140 x 210 mm (khổ A4 gập đôi). Tóm tắt luận văn được trình bày nhiều nhất trong 24 trang in trên hai mặt giấy, cỡ chữ Times New Roman 11 của hệ soạn thảo Winword hoặc phần mềm soạn thảo Latex đối với các chuyên ngành thuộc ngành Toán. Mật độ chữ bình thường, không được nén hoặc kéo dãn khoảng cách giữa các chữ. Chế độ dãn dòng là Exactly 17pt. Lề trên, lề dưới, lề trái, lề phải đều là 1.5 cm. Các bảng biểu trình bày theo chiều ngang khổ giấy thì đầu bảng là lề trái của trang. Tóm tắt luận án phải phản ảnh trung thực kết cấu, bố cục và nội dung của luận án, phải ghi đầy đủ toàn văn kết luận của luận án. Mẫu trình bày trang bìa của tóm tắt luận văn (phụ lục 1).

Chương 2

Tổng quan về tóm tắt văn bản

2.1 Giới thiệu về tóm tắt văn bản

2.1.1 Giới thiệu chung

Từ công trình đầu tiên của Luhn năm 1958, tóm tắt văn bản đã và đang trở thành một trong những tác vụ phổ biến và cần thiết nhất. Đặc biệt, trong sự bùng nổ không ngừng về công nghệ thông tin, việc tiếp nhận và xử lý khối lượng thông tin ngày càng lớn đang trở thành bài toán vô cùng thiết thực và quan trọng. Theo Luhn, mục đích của tóm tắt là nhằm tạo điều kiện giúp xác định nhanh chóng và chính xác chủ đề của văn bản gốc. Mục tiêu là tiết kiệm thời gian và công sức của người đọc trong việc tìm kiếm thông tin hữu ích của văn bản hoặc báo cáo.

2.1.2 Phân loại tóm tắt văn bản

Bài toán tóm tắt văn bản được phân loại dựa trên nhiều khía cạnh khác nhau. Mỗi khía cạnh được áp dụng cho một mục đích khác nhau cũng

như đòi hỏi các cách giải quyết khác nhau. Do đó, khó có một phương pháp chung nào có thể áp dụng tổng quát cho tất cả các loại tóm tắt văn bản. Vì vậy, cần xác định rõ đối tượng cũng như mục tiêu bài toán để chọn lựa phương pháp giải quyết cho phù hợp. Nhìn chung, tóm tắt văn bản có thể được chia theo một số dạng sau:

- Về hình thức, tóm tắt văn bản được chia làm 2 loại: tóm tắt tóm lượt (abstractive) và tóm tắt rút trích (extractive). Theo đó, tóm tắt rút trích được tạo ra bằng các nối kết các câu được trích nguyên gốc từ văn bản ban đầu. Trong khi đó, tóm tắt tóm lượt được tạo ra bằng ngôn ngữ của người tóm tắt dựa trên nội dung của văn bản ban đầu.

- Về đối tượng, tóm tắt văn bản được chia làm 2 loại: đơn văn bản (single document) và đa văn bản (multi documents). Với tóm tắt đơn văn bản, đầu vào của bài toán chỉ là một văn bản xác định. Khác với nó, tóm tắt đa văn bản nhận đầu vào là một tập các văn bản khác nhau mà có cùng chủ đề hoặc sự kiện.

- Về nội dung, tóm tắt văn bản được chia làm tóm tắt chỉ định (indicative) và tóm tắt thông tin (informative). Mục đích của tóm tắt chỉ định là giúp người đọc quyết định xem có nên tiếp tục đọc hay không bằng việc cung cấp các đặc trưng của văn bản như: chiều dài, văn phong, ... Trong khi đó, tóm tắt thông tin cung cấp các sự kiện, vấn đề được tường thuật trong văn bản đầu vào.

- Về mục đích, tóm tắt văn bản được chia làm: tóm tắt tổng quát (generic) và tóm tắt hướng truy vấn (query focused). Tóm tắt tổng quát đặt ra giả thuyết độc giả là chung chung. Trong khi đó, mục đích của tóm tắt hướng truy vấn là tóm tắt các thông tin liên quan dựa trên một số yêu cầu truy vấn của người dùng.

2.2 Các phương pháp tóm tắt văn bản

2.3 Các vấn đề trong tóm tắt văn bản

2.3.1 Sắp xếp câu (Sentence ordering)

Trong tóm tắt trích xuất, các câu quan trọng từ văn bản đầu vào sẽ được chọn lọc và đưa vào văn bản tóm tắt đầu ra. Tuy nhiên, văn bản là một chỉnh thể thống nhất có thứ tự của các câu văn. Vì thế, những câu quan trọng sau khi được chọn lọc cần phải trải qua việc tái sắp xếp để đảm bảo tính đúng đắn về mặt ngữ nghĩa. Một cách trực quan, sắp xếp câu nghĩa là tìm ra một trật tự có nghĩa của một tập câu cho trước sao cho văn bản tóm tắt phản ánh gần nhất văn bản đầu vào.

2.3.2 Giản lược câu (Sentence revision)

Giản lược câu là việc biến đổi văn bản tóm tắt tạm thời bằng cách thay thế hoặc điều chỉnh các từ hoặc ngữ bằng từ hoặc ngữ khác thích hợp hơn dựa trên ngữ cảnh của văn bản tóm tắt. Về cơ bản, một số loại giản lược được đề xuất gồm : loại bỏ những thành phần (câu, ngữ, từ) không cần thiết, kết hợp các thông tin từ những câu khác nhau và thay đổi thành phần này bằng thành phần khác. Tuy nhiên, tác vụ này tương đối phức tạp và thường được nghiên cứu tập trung vào từng loại giản lược khác nhau.

2.3.3 Kết hợp câu

Hợp nhất câu (Sentence fusion)

Hợp nhất câu là tác vụ được thực hiện trên hai câu mà trong đó có sự trùng lắp một số thông tin. Mục đích của nó là tạo ra một câu mới chứa

thông tin trùng lắp của các câu đầu vào.

Nén câu (Sentence compression)

Nhiều nhà nghiên cứu nhận thấy các văn bản tóm tắt thường chứa chỉ một phần của văn bản gốc. Những thành phần, yếu tố không cần thiết trong câu như bổ ngữ, mệnh đề phụ, chú giải, ... thường được lượt bỏ nhằm giúp tăng tính xúc tích của văn bản tóm tắt.

Chương 3

Tổng quan về Deep Learning

3.1 Feed-forward Neural Network

3.2 Recursive Neural Network

3.2.1 Recursive Neural Network

3.2.2 Recursive Auto-encoder Neural Network

3.3 Convolution Neural Network

3.4 Recurrent Neural Network

3.4.1 Recurrent Neural Network

3.4.2 Long Short Term Memory

Tài liệu tham khảo

Tiếng Anh

[1] Cavnar, William B. and Trenkle, John M. "N-Gram-Based Text Categorization". In: *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*. 1994, pp. 161–175.

[2] Knuth, Donald E. *The TEXbook*. Addison-Wesley, 1984.

[3] Online. *LaTeX/Floats, Figures and Captions*. URL: `http://en.wikibooks.org/wiki/LaTeX/Floats,_Figures_and_Captions` (visited on 06/06/2015).

[4] Online. *LaTeX/Source Code Listings*. URL: `http://en.wikibooks.org/wiki/LaTeX/Source_Code_Listings` (visited on 06/06/2015).

[5] Online. *LaTeX/Tables*. URL: `http://en.wikibooks.org/wiki/LaTeX/Tables` (visited on 06/06/2015).

[6] Zhang, Kaizhong and Shasha, Dennis. "Simple fast algorithms for the editing distance between trees and related problems". In: *SIAM Journal on Computing, Volume 18 Issue 6* (1989), pp. 1245–1262.

Phụ lục A

Ngữ pháp tiếng Việt

Đây là phụ lục.

Phụ lục B

Ngữ pháp tiếng Nôm

Đây là phụ lục 2.