BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

LÊ MINH HIẾU

PHÂN ĐOẠN TỪ TIẾNG VIỆT

LUẬN VĂN THẠC SĨ

Chuyên ngành: Công nghệ thông tin

Mã số ngành:60480201

TP. HỒ CHÍ MINH, tháng 01 năm 2015

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

LÊ MINH HIẾU

PHÂN ĐOẠN TỪ TIẾNG VIỆT

LUẬN VĂN THẠC SĨ

Chuyên ngành: Công nghệ thông tin

Mã số ngành:60480201

CÁN BỘ HƯỚNG DẪN KHOA HỌC:TS.NGUYỄN CHÍ HIẾU

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

Cán bộ hướng dẫn khoa học: **TS.NGUYỄN CHÍ HIẾU**
(Ghi rõ họ, tên, học hàm, học vị và chữ ký)

TS. Nguyễn Chí Hiếu

Luận văn Thạc sĩ được bảo vệ tại Trường Đại học Công nghệ TP. HCM ngày06 tháng 02năm2015

Thành phần Hội đồng đánh giá Luận văn Thạc sĩ gồm:
(Ghi rõ họ, tên, học hàm, học vị của Hội đồng chấm bảo vệ Luận văn Thạc sĩ)

TT	**Họ và tên**	**Chức danh Hội đồng**
1	PGS.TS. Lê Hoài Bắc	Chủ tịch
2	PGS.TS. Quản Thành Thơ	Phản biện 1
3	TS. Võ Đình Bảy	Phản biện 2
4	TS. Lư Nhật Vinh	Ủy viên
5	TS. Cao Tùng Anh	Ủy viên, Thư ký

Xác nhận của Chủ tịch Hội đồng đánh giá Luận vănsau khi Luận văn đã được sửa chữa (nếu có).

Chủ tịch Hội đồng đánh giá LV

PGS.TS. Lê Hoài Bắc

TRƯỜNG ĐH CÔNG NGHỆ TP. HCM
PHÒNG QLKH – ĐTSĐH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

TP. HCM, ngày… tháng ... năm 20…

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên: **LÊ MINH HIẾU**Giới tính:**NAM**

Ngày, tháng, năm sinh: **20/10/1985**Nơi sinh:**GIA LAI**

Chuyên ngành: **CÔNG NGHỆ THÔNG TIN**MSHV:**1241860004**

I- Tên đề tài: Phân đoạn từ tiếng Việt

II- Nhiệm vụ và nội dung:

- Nghiên cứu cơ sở lý thuyết về xử lý ngôn ngữ tự nhiên.
- Khảo sát các nghiên cứu liên quan.
- Xây dựng mô hình phân đoạn từ tiếng Việt.
- Chạy thực nghiệm và đánh giá kết quả.

III- Ngày giao nhiệm vụ:*(Ngày bắt đầu thực hiện LV ghi trong QĐ giao đề tài)*

.....

IV- Ngày hoàn thành nhiệm vụ:(Ngày bảo vệ LV)

.....

V- Cán bộ hướng dẫn: TS.NGUYỄN CHÍ HIẾU

CÁN BỘ HƯỚNG DẪN **KHOA QUẢN LÝ CHUYÊN NGÀNH**

(Họ tên và chữ ký) (Họ tên và chữ ký)

TS. Nguyễn Chí Hiếu

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong Luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện Luận văn này đã được cảm ơn và các thông tin trích dẫn trong Luận văn đã được chỉ rõ nguồn gốc.

Học viên thực hiện Luận văn

(Ký và ghi rõ họ tên)

Lê Minh Hiếu

LỜI CẢM ƠN

Với tất cả tấm lòng, tôi xin gửi lời cảm ơn sâu sắc nhất đến thầy giáo TS.Nguyễn Chí Hiếu – người thầy đã tận tình hướng dẫn, chỉ bảo và tạo những điều kiện tốt nhất giúp tôi hoàn thành luận văn này.

Đồng thời tôi xin gửi lời cảm ơn chân thành đến toàn thể quý thầy cô trường Đại học Công nghệ Thành phố Hồ Chí Minh đã trang bị cho tôi những kiến thức trong học tập và nghiên cứu khoa học.

Tôi cũng xin chân thành cảm ơn các thành viên trong đề tài "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt", mã số KC01.01/06-10 đã cho phép tôi sử dụng một số dữ liệu của VietTreebank và Vietnamese Lexicon trong quá trình thực nghiệm.

Cuối cùng, tôi xin gửi lời cảm ơn đến gia đình, bạn bè và các đồng nghiệp đã luôn động viên và cho tôi những lời khuyên bổ ích trong suốt quá trình thực hiện luận văn này.

Tp.Hồ Chí Minh, tháng 01 năm 2015

Lê Minh Hiếu

TÓM TẮT

Từ khóa

Phân đoạn từ, phân giải nhập nhằng, nhận dạng danh từ riêng, thông tin tương hỗ.

Tóm tắt

Không giống như tiếng Anh, phân đoạn từ trong ngôn ngữ tiếng Việt, cũng như hầu hết các ngôn ngữ ở châu Á, là một công việc hết sức phức tạp.Vì bản thân ngôn ngữ không có những dấu hiệu rõ ràng để phân cách các từ với nhau, chẳng hạn như khoảng trắng. Đã có rất nhiều nghiên cứu với nhiều hướng tiếp cận khác nhau về công việc này. Tuy nhiên theo khảo sát, đại đa số các nghiên cứu đều xuất phát từ ba hướng tiếp cận chính: hướng tiếp cận dựa trên từ điển, hướng tiếp cận dựa trên mô hình thống kê và hướng tiếp cận lai.

Rất nhiều nghiên cứu đã chọn hướng tiếp cận dựa trên từ điển vì tính đơn giản của nó. Hướng tiếp cận này thường sử dụng từ điển kết hợp với một số thuật toán so khớp như: Maximum matching (MM), Longest matching (LM), v.v... để phân đoạn từ. Tuy nhiên hướng tiếp cận này thường gây ra nhiều nhập nhằng khi phân đoạn và không thể phân đoạn đúng cho các từ không có trong từ điển.

Hướng tiếp cận dựa trên thống kê cần một kho ngữ liệu đủ lớn, đã trải qua quá trình huấn luyện, kết hợp với các thuật toán thống kê để phân đoạn từ. Có thể kể đến một số mô hình theo hướng tiếp cận này như: mô hình thống kê N-gram, mô hình cực đại hóa Entropy (ME), mô hình Conditional Random Fields (CRFs), mô hình cây quyết định. Ưu điểm của hướng tiếp cận này là có thể phát hiện được các từ không có trong từ điển và hạn chế được nhập nhằng.

Hướng tiếp cận dựa trên mô hình lai kết hợp nhiều phương pháp khác nhau để phân đoạn từ. Một số mô hình phân đoạn từ theo mô hình lai có thể kể đến như: mô hình so khớp Maximum matching kết hợp với SVMs, mô hình phân đoạn từ sử dụng WFST và mạng Neural, mô hình sử dụng thuật toán Maximum matching và N-gram, mô hình kết hợp CRFs và SVMs. Hướng tiếp cận này thường phức tạp nhưng mang lại hiệu quả cao.

Trong luận văn này chúng tôi đề xuất một mô hình phân đoạn từ dựa trên mô hình lai. Mô hình của chúng tôi sử dụng bốn luật phân giải nhập nhằng của hệ thống MMSeg kết hợp vớitừ điển, thông tin huấn luyện N-gram, thông tin hỗ tươngvà các biểu thức chính quy.

Thực nghiệm trên văn bản gồm 10,000 câu trích từVietTreebank cho kết quả F-measure đạt 91.74%.

ABSTRACT

Keywords

Vietnamese word segment, disambiguity, proper nouns identification, mutual information.

Abstract

Unlike in English, word segmentation in Vietnamese, as well as in many other Asian languages, is more complex because the language does not have any explicit word boundary delimiters, such as a space, to separate between each word. Many researchers with many approaches for the word segmentation task. However, these approaches can be classified into 3 major categories: dictionary-based, statistics-based and hybrid-based.

Most studies use dictionary-based approaches because of their simplicity. This approach type use dictionaries with matching methods as Maximum matching (MM), Longest matching (LM), ect for the word segmentation. However, most of the dictionary based approaches often get many ambiguous cases and can not detects new words.

Statistical approaches need a very large annotated training corpus for word segmentation. Some of studies based on this approaches are N-gram Language Model, Maximum Entropy (ME),Conditional Random Fields (CRFs), Decision Tree. This approach is usefull for detects new words and disambiguous.

Hybrid approaches combine different approaches to make use of individual advantages and overcome disadvantages. Some models are combination of Maximum matching and SVMs, WFST and Neural network, Maximum matching and Ngram language model, CRFs and SVMs. This approache are often complex however it give a high accuracy.

In this thesis, we propose a hybrid method for Vietnamese word segmentation. Our approach is base on four MMSegdisambiguity rules, dictionaries, ngram trained corpus, mutual information and regular expression.

Experiment on 10,000 sentences of VietTreebank corpus gives a result with an F-measure of 91.74%.

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

STT	Từ viết tắt	Diễn giải tiếng Anh	Diễn giải tiếng Việt
1	ATS	Automatic Text Summarization	Tóm lược văn bản
2	CRFs	Conditional Random Fields	Học máy CRFs
3	DCB	Dictionary Based	Dựa trên từ điển
4	DT	Decision Tree	Cây quyết định
5	HMM	Hidden Markov Model	Mô hình Markov ẩn
6	IR-IE	Information Retrieval and Extraction	Truy vấn và khai thác thông tin
7	LM	Longest Matching	So khớp dài nhất
8	ME	Maximum Entropy	Cực đại hóa Entroy
9	MLB	Machine Learning Based	Dựa trên học máy
10	MM	Maximal Matching	So khớp cực đại
11	MT	Machine Translation	Dịch máy
12	Q&A	Question and Answer	Hệ thống hỏi đáp
13	SR	Speech Recognition	Nhận dạng tiếng nói
14	SVMs	SVMs	Máy học vectơ hỗ trợ
15	WS	Word Segmention	Phân đoạn từ

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ

GIỚI THIỆU

1. Đặt vấn đề

Xử lý ngôn ngữ tự nhiên (NLP: Natural Language Processing) là một nhánh của trí tuệ nhân tạo, tập trung vào các ứng dụng trên ngôn ngữ con người. Xử lý ngôn ngữ tự nhiên góp phần trong việc làm cho máy móc có thể hiểu được ngôn ngữ con người, từ đó tạo ra các hệ thống thông minh.

Nghiên cứu về xử lý ngôn ngữ tự nhiên bao gồm nhiều lĩnh vực quan trọng như: dịch máy (MT: Machine Translation), truy vấn và khai thác thông tin (IR-IE: Information Retrieval and Extraction), hệ thống hỏi đáp (Q&A: Question and Answer), tóm lược văn bản (ATS: Automatic Text Summarization), nhận dạng tiếng nói (SR: Speech Recognition), v.v... từng bước giúp máy tính hiểu được con người.

Phân đoạn từ (WS: Word Segmention) là một bước quan trọng trong xử lý ngôn ngữ tự nhiên tiếng Việt, đặc biệt là xử lý văn bản. Phân đoạn từ là việc xác định ranh giới giữa các từ trong câu.

Không giống như tiếng Anh và các ngôn ngữ Ấn-Âu khác, tiếng Việt không sử dụng khoảng cách làm dấu hiệu xác định ranh giới từ. Ranh giới giữa các từ không có dấu hiệu rõ ràng mà cần phải dựa vào các yếu tố như: ngữ nghĩa, ngữ cảnh, văn phong, các từ lân cận, v.v...

Ngoài ra, vấn đề từ đa nghĩa, từ ghép cũng gây nhiều khó khăn trong việc phân đoạn từ tiếng Việt. Phân đoạn từ có độ chính xác cao sẽ góp phần quan trọng vào các bài toán tiếp theo như: gán nhãn từ loại, kiểm tra cú pháp, dịch tự động, v.v...

2. Lý do chọn đề tài

Với các ngôn ngữ biến hình như tiếng Anh, Pháp, Đức, Nga, … việc nhận biết ranh giới giữa các từ đơn giản hơn tiếng Việt, chủ yếu dựa vào khoảng cách và các dấu câu. Bản thân các từ hầu như đã phản ánh đầy đủ hình thái, nghĩa, thậm chí ngữ pháp bên trong nó.

Tuy nhiên, tiếng Việt là ngôn ngữ thuộc hệ đơn lập, không biến hình. Về mặt hình thức một từ có thể được cấu tạo bởi một hoặc nhiều âm tiết ghép lại. Khoảng trắng chỉ dùng để phân cách các âm tiết với nhau. Để có thể tiến tới các xử lý xa hơn về xử lý ngôn ngữ tự nhiên trước hết ta phải làm tốt bài toán phân đoạn từ. Từ là đơn vị cơ bản nhất để phân tích cú pháp, ngữ nghĩa của ngôn ngữ.

Cho đến nay, đã có rất nhiều công trình nghiên cứu về phân đoạn từ tiếng Việt với những kết quả khả quan. Tuy nhiên các vấn đề như: hiện tượng phát sinh từ mới, sự nhập nhằng ngữ nghĩa, v.v...đã ảnh hưởng không ít đến chất lượng phân đoạn từ. Vì vậy phân đoạn từ tiếng Việt vẫn là chủ đề được nhiều nhà nghiên cứu quan tâm và là động lực của luận văn này.

3. Mục tiêu và phạm vinghiên cứu

Chúng tôi đặt ra mục tiêu nghiên cứu chính của luận văn là xây dựng một mô hình phân đoạn từ tiếng Việt dựa trên mô hình lai kết hợp nhiều phương pháp nhằm tăng cường độ chính xác khi phân đoạn từ.

Phạm vi của đề tài tập trung nghiên cứu phân đoạn từ trên văn bản tiếng Việt. Với đầu vào là một văn bản tiếng Việt, đầu ra là một văn bản tiếng Việt đã được phân đoạn thành các từ.

Với mục tiêu nêu trên, luận văn tập trung nghiên cứu các vấn đề sau đây:

- ❖ Nghiên cứu tổng quan về xử lý ngôn ngữ tự nhiên. Khảo sát các công trình nghiên cứu có liên quan đến đề tài trong nước và quốc tế.

- ❖ Nghiên cứu cơ sở lý thuyết về ngôn ngữ bao gồm: các loại hình ngôn ngữ, đơn vị chủ yếu của ngôn ngữ tiếng Việt, cấu trúc của đơn vị từ trong tiếng Việt, nghiên cứu về từ vựng và hiện tượng nhập nhằng nghĩa của từ.
- ❖ Nghiên cứu cơ sở lý thuyết về ngôn ngữ học thống kê bao gồm: lý thuyết xác suất thống kê trong xử lý ngôn ngữ tự nhiên, mô hình Markov ẩn, mô hình thống kê N-Gram.
- ❖ Xây dựng kho ngữ liệu phục vụ các mô hình thống kê.
- ❖ Thu thập và xây dựng từ điển từ vựng, từ điểndanh từ riêng.
- ❖ Nghiên cứu các phương pháp phân đoạn từ dựa trên từ điển.
- ❖ Nghiên cứu các phương pháp phân đoạn từ dựa trên mô hình thống kê.
- ❖ Nghiên cứu cácphương pháp phát hiện từ mới sử dụng kho ngữ liệu và thông tin tương hỗ.
- ❖ Xây dựng mô hình phân đoạn từ tiếng Việt bằng cách kết hợp các phương pháp: phương pháp phân đoạn từ có tham khảo từ điển từ vựng, phương pháp nhận dạng danh từ riêng sử dụng từ điển danh từ riêng, phương pháp so trùng các mẫu dùng biểu thức chính quy và phương pháp tự động phát hiện từ mới sử dụng thông tin tương hỗ.

4. Bố cục của luận văn

Luận văn được tổ chức gồm có 5 chương. Chương 1: trình bày tổng quan về các hướng tiếp cận và các công trình nghiên cứu có liên quan đến đề tài. Chương 2: trình bày về cơ sở lý thuyết của đề tài, bao gồm cơ sở lý thuyết về ngôn ngữ và ngôn ngữ học thống kê. Chương 3: giới thiệu mô hình MMSeg – mô hình tham khảo chính của đề tài. Chương 4: giới thiệu mô hình phân đoạn từ do luận văn đề xuất. Chương 5: kết luận, tự đánh giá và nhận xét về những kết quả đạt được, những mặt còn hạn chế và hướng phát triển của đề tài.

CHƯƠNG 1. TỔNG QUAN

Không giống như tiếng Anh và các ngôn ngữ Ấn-Âu sử dụng khoảng cách làm dấu hiệu phân cách từ, hầu hết các ngôn ngữ châu Á (như tiếng Việt, tiếng Thái, tiếng Nhật, v.v...) phải dựa vào nhiều yếu tố (như ngữ nghĩa, ngữ cảnh, các từ lân cận, v.v...) mới có thể xác định được ranh giới giữa các từ. Cho đến nay đã có rấtnhiều công trình nghiên cứu về phân đoạn từ với nhiều phương pháp khác nhau. Theo khảo sát của chúng tôi các nghiên cứuhầu hết xuất phát từ 3 hướng tiếp cận chính sau đây: hướng tiếp cận dựa trên từ điển (dictionary-based), hướng tiếp cận dựa trên thống kê (statistics-based) và hướng tiếp cận lai (hybrid-based).

Hướng tiếp cận dựa trên từ điển: đây là hướng tiếp cận cơ bản nhất. Đặc điểm chung của hướng tiếp cận này là sử dụng từ điển từ vựng kết hợp với cácthuật toán so khớp để phân đoạn từ. Độ chính xác của phân đoạn phụ thuộc vào tính đầy đủ của từ điển. Hướng tiếp cận này có ưu điểm: tốc độ xử lý nhanh, đơn giản. Tuy nhiên có hạn chếlà không thể xác định được các từ không có trong từ điển, nhập nhằng phân đoạn có thể xảy ra lớn.

Hướng tiếp cận dựa trênthống kê hoặc thống kê kết hợp với học máy: hướng tiếp cận nàycó đặc điểm cần phải xây dựng kho ngữ liệu bằng cách thu thập dữ liệu về ngôn ngữ, sau đó tiến hànhthống kê, học máy trên kho ngữ liệu thu thập được (gọi là huấn luyện dữ liệu), dựa trên dữ liệu huấn luyện và các thuật toán để phân đoạn từ. Độ chính xác của phương pháp phụ thuộc nhiều vào độ lớn và độ bao quát của kho ngữ liệu. Ưu điểm của hướng tiếp cận này là có thể phân đoạn được các từ mới, hạn chế được nhập nhằng phân đoạn nhưng có hạn chế là tốnnhiều thời gian, công sức để xây dựng và xử lý kho ngữ liệu.

Hướng tiếp cận lai: sử dụng kết hợp cùng lúc nhiều phương pháp để tăng cường độ chính xác của phân đoạn. Ưu điểm: độ chính xác được tăng cường. Nhược điểm: độ phức tạp lớn.

Trong phần tiếp theo, chúng tôi nêu kết quả khảo sát và mô tả một số công trình nghiên cứu có liên quan đến đề tài. Các nghiên cứu này được thực hiện trên ngôn ngữ tiếng Việt hoặc trên những ngôn ngữ có đặc điểm tương đồng với tiếng Việt.

Trên ngôn ngữ tiếng Myanmar, Hla Hla Htay và Kavi Narayana Murthy trong [14] sử dụng thuật toán so khớp dài nhất (LM: Longest Matching) để phân đoạn từ tiếng Myanma. Từ điển được xây dựng bằng cách tập hợp khoảng 4550 âm tiết có trong ngôn ngữ, sau đó tiến hành gộp âm tiết để tạo nên khoảng 800,000 từ và các biến thể của từ. Thực nghiệm được tiến hành trên 5000 câu (chứa 35049 từ). Kết quả thu được 34,943 từ với 34,633 từ đúng. Độ chính xác F-measuređạt 98.95%.

Trên ngôn ngữ tiếng Hoa, Jin Kiat Low và cộng sự trong [18] sử dụng mô hình cực đại Entropy (ME: Maximum Entropy) có tham khảo từ điển để phân đoạn từ tiếng Trung Quốc. Từ điển được sử dụng chứa khoảng 108.000 từ. Thực nghiệm được tiến hành đồng thời trên bốn corpus khác nhau: Academia Sinica (AS), City University of Hong Kong (CITYU), Microsoft Research (MSR) và Peking University (PKU). Kết quả F-measure đạt từ 95,6% - 96,9%.

Trên ngôn ngữ tiếng Nhật, Masaaki Nagata trong [20]đề xuất một mô hình phân đoạn từ tiếng Nhật dựa trên thống kê. Ở bước khởi tạo, mô hình sử dụng một tập hợp các từ cơ bản gọi là *word base*. Sau đó, tiến hành huấn luyện kho ngữ liệu dựa trên việc tính toán tần số xuất hiện của các chuỗi trong tập ngữ liệu. Tiếp theo, word base được tăng cường bởi các từ xác định được trong quá trình huấn luyện. Cuối cùng, phương pháp thực hiện đánh giá lại để loại bỏ những từ không phù hợp trong word base. Khi kho ngữ liệu đạt đến độ lớn 3.9Mb với khoảng 1791 từ cơ bản, độ chính xác accuracy của phương pháp đã đạt 82,5%. Phương pháp này sử dụng word base như làm kinh nghiệm để phân đoạn và không cần word base có kích thước lớn ở giai đoạn ban đầu.

Trên ngôn ngữ tiếng Thái, Thanaruk Theeramunkong và Sasiporn Usanavasin trong [24]xây dựng mô hình phân đoạn từ tiếng Thái dựa trên cây quyết định không dùng từ điển. Sử dụng cấu trúc từ vựng tiếng Thái làm dấu hiệu đặc trưng để phân lớp. Ở giai đoạn huấn luyện, tác giả tạo một copus nhằm xây dựng cây quyết định. Sau đó văn bản tiếng Thái sẽ được phân đoạn dựa trên luật của cây quyết định. Luật của cây quyết định được xây dựng dựa vào những kí tự nằm kề nhau không thể tách

rời, gọi là “Thai character clusters - TCCs”. Thực nghiệm trên kho ngữ liệu tiếng Thái, kết quả độ chính xác accuracy đạt 87.41%.

Trên ngôn ngữ tiếng Việt, nhiều mô hình phân đoạn từ đã được nghiên cứu và đề xuất với những kết quả khả quan. Lê Trung Hiếu và cộng sự trong [13]xây dựng mô hình xác suất nhận dạng và phân tách từ tiếng Việt, đồng thời áp dụng quá trình máy tự học xây dựng mô hình xác suất tối ưu. Độ chính xác của thuật toán phân tách từ đạt trên 90%.

Trần Ngọc Anh và cộng sự trong [3] đề xuất một phương pháp phân đoạn từ và xử lý nhập nhằng phân đoạn dựa trên mô hình lai. Sử dụng kỹ thuật so khớp cực đại (MM: Maximum Matching) để phân đoạn từ. Trong quá trình phân đoạn, tác giả sử dụng đồng thời phương pháp (FMM: Foward Maximum Matching) và (BMM: Backward Maximum Matching) nhằm phát hiện nhập nhằng. Sau đó xử lý nhập nhằng bằng cách kết hợp nhiều phương pháp, bao gồm: phương pháp thống kê dựa trên mô hình Bi-Gram trên từ, mô hình N-Gram dựa trên âm tiết, và phương pháp tham khảo từ điển. Thực nghiệm trên corpus đã được huấn luyện với 2639 tập tin văn bản, với 1,541,188 từ. Kết quả độ chính xác F-measure đạt 98.71% - 98.94%.

Lưu Tuấn Anh và Yamamoto Kazuhide trong [2] xây dựng mô hình phân đoạn từ với hướng tiếp cận Pointwise dựa trên máy học SVM. Kết quả của nghiên cứu được ứng dụng xây dựng công cụ tách từ có tên là Đông Du với độ chính xác 98,2 %.

Lê Hồng Phương và cộng sự trong [15]sử dụng mô hình lai dựa trên kĩ thuật so khớp cực đại kết hợp automat hữu hạn trạng thái và regular expression. Ngoài ra, để xử lí nhập nhằng, hệ thống kết hợp với các thống kê Uni-Gram và Bi-Gram huấn luyện trên tập văn bản tách từ mẫu. Kết quả nghiên cứu được ứng dụng tạo nên công cụ vnTokenizer với độ chính xác F-measure đạt được gần 94%.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1 Cơ sở lý thuyết về ngôn ngữ

2.1.1 Phân loại ngôn ngữ

Xét theo loại hình ngôn ngữ, theo Nguyễn Thiện Giáp trong [9, tr 298–305]ngôn ngữ có thể chia làm 2 loại chính: ngôn ngữ đơn lập và ngôn ngữ không đơn lập. Đơn lập có thể hiểu theo hai cách: đơn lập về ngữ âm và đơn lập về ngữ pháp. Đơn lập về ngữ âm giống như tính đơn tiết của từ hay hình vị. Đơn lập về ngữ pháp nói đến tính độc lập của từ hoạt động trong câu. Sự khác biệt cơ bản giữa 2 loại hình này là đặc điểm cấu tạo của từ.

2.1.1.1 Ngôn ngữ không đơn lập

Ngôn ngữ không đơn lập được chia làm 3 loại chính: ngôn ngữ chắp dính, ngôn ngữ hòa kết và ngôn ngữ hỗn nhập.

- **Ngôn ngữ chắp dính**

Đặc điểm của loại ngôn ngữ này là sử dụng rộng rãi các phụ tố để cấu tạo từ và biểu thị những mối quan hệ khác nhau. Mỗi phụ tố chỉ biểu thị cho một ý nghĩa ngữ pháp và ngược lại. Hình vị trong các ngôn ngữ chắp dính có tính độc lập lớn và mối liên hệ giữa các hình vị không chặt chẽ. Chính tố có thể hoạt động độc lập.

Ví dụ, trong tiếng Thổ Nhĩ Kì:

- adam: người đàn ông

- adamlar: những người đàn ông

- kadin: người đàn bà:

- kadinlar: những người đàn bà

Có thể liệt kê một số ngôn ngữ thuộc loại này như: tiếng Thổ Nhĩ Kỳ, tiếng Ugo-Phần Lan, tiếng Bantu, v.v…

❖ **Ngôn ngữ hòa kết**

Còn được gọi là ngôn ngữ chuyển dạng. Đặc điểm của loại hình ngôn ngữ này là có sự biến đổi giữa nguyên âm và phụ âm trong hình vị mang ý nghĩa ngữ pháp. Ý nghĩa từ vựng và ý nghĩa ngữ pháp được dung hợp ở trong từ nhưng không thể tách bạch phần nào biểu thị ý nghĩa từ vựng, phần nào biểu thị ý nghĩa ngữ pháp. Mỗi phụ tố có thể mang đồng thời nhiều ý nghĩa và ngược lại. Các hình vị liên kết chặt chẽ với nhau.

Ví dụ:

Bảng 2.1 Bảng minh họa ngôn ngữ hòa kết

Tiếng Anh	foot: bàn chân– feet: những bàn chân
Tiếng Ả Rập	balad: làng– biläd: những làng

Các ngôn ngữ chuyển dạng gồm các tiếng Ấn-Âu hiện đại như các tiếng Pháp, tiếng Ý, tiếng Anh, tiếng Bungari, v.v…

❖ **Ngôn ngữ hỗn nhập**

Đặc điểm của các ngôn ngữ hỗn nhập là một từ có thể tương ứng với một câu trong các ngôn ngữ khác. Nghĩa là đối tượng hành động, trạng thái hành động không được thể hiện bằng các thành phần câu đặc biệt như tân ngữ, trạng ngữ, định ngữ v.v... mà được thể hiện bằng các phụ tố khác nhau trong hình thái động từ.

Ví dụ: trong tiếng Tschinuk ở Bắc Mĩ, từ **“inialudam”** tương ứng với câu **"Tôi đã đến để cho cô cái này"**.

Bảng 2.2 Bảng minh họa một từ trong tiếng Tschinuk

Ký tự	i	n	i	a	l	u	d	a	m
Vị trí	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]

Trong đó:

- Phụ âm d[7] là động từ chính trong câu, có nghĩa là ***cho***
- Tiền tố i[1] biểu hiện thì quá khứ, có nghĩa là ***đã***
- Phụ tố n[2] biểu hiện ngôi thứ nhất số ít, có nghĩa là ***tôi***
- Phụ tố i[3] biểu hiện tân ngữ giới từ, có nghĩa là***cái này***
- Phụ tố a[4] biểu hiện tân ngữ của giới từ, có nghĩa là ***cô***
- Phụ tố l[5] cho biết tân ngữ của giới từ ***cô*** là gián tiếp
- Phụ tố u[6] chỉ ra rằng hành động ở ***thể chủ động***
- Phụ tố am[8,9] nhấn mạnh **tính có mục đích** của hành động.

Một số ngôn ngữ Ấn ở Nam Mĩ và Đông Nam Xibêri v.v... cũng thuộc loại ngôn ngữ hỗn nhập.

2.1.1.2 Ngôn ngữ đơn lập

Về mặt ngữ pháp, từ trong ngôn ngữ đơn lập không biến đổi hình thái. Cấu tạo từ do căn tố hoặc sự kết hợp giữa các căn tố tạo thành. Quan hệ ngữ pháp và ý nghĩa ngữ pháp được thể hiện bằng các phương tiện ngoài từ như: trật tự từ, hư từ, ngữ điệu, v.v...

Ví dụ:

Dùng hư từ
- Cuốn sách – những cuốn sách
- Đi – sẽ đi, đang đi, đã đi, mới đi, …

Dùng trật tự từ
- Cửa trước – trước cửa
- Nhà nước – nước nhà
- Xanh mặt – mặt xanh

Ranh giới giữa từ ghép và cụm từ đôi khi khó phân biệt rõ ràng. Ví dụ: xe đạp, nhà phố, v.v…

Về mặt ngữ âm, ngôn ngữ đơn lập thể hiện rõ mối quan hệ giữa hình vị và âm tiết. Ranh giới giữa hình vị trùng với âm tiết tạo nên hình tiết. Hình tiết là một đơn

vị có vỏ ngữ âm là âm tiết, có khi được dùng với tư cách một từ, có khi được dùng với tư cách là yếu tố cấu tạo từ.

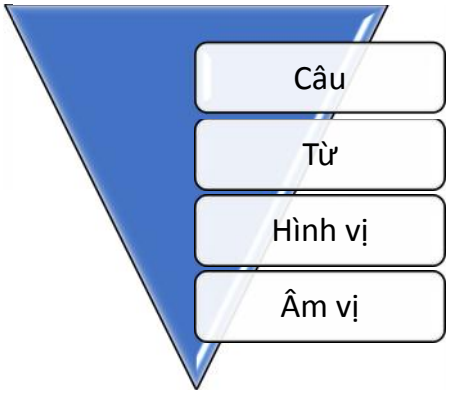
Âm tiết trong ngôn ngữ đơn lập có cấu trúc chặt chẽ. Mỗi âm vị nằm ở vị trí nhất định, có chức năng nhất định.

Tiêu biểu cho ngôn ngữ đơn lập có thể kể đến một số ngôn ngữ như: tiếng Hán, tiếng Thái, tiếng Dao, tiếng Miến Điện, tiếng Khmer, v.v...

Tiếng Việt cũng thuộc loại hình ngôn ngữ đơn lập, không biến đổi hình thái. Số lượng vỏ âm thanh mà người Việt sử dụng làm vỏ ngữ âm cho hình vị tối đa khoảng 4 vạn tiếng khác nhau. [5, tr.46]

2.1.2 Đơn vị chủ yếu của ngôn ngữ

Theo Nguyễn Thiện Giáp trong [11, tr 52-55] đơn vị chủ yếu của ngôn ngữ gồm có:



Hình 2.1 Hình minh họa các đơn vị chủ yếu của ngôn ngữ

2.1.2.1 Âm vị

Âm vị còn được gọi là âm tiết là đơn vị tối thiểu của hệ thống ngữ âm của một ngôn ngữ dùng để cấu tạo và phân biệt vỏ âm thanh của các đơn vị có nghĩa của ngôn ngữ.

Tiếng Việt thuộc loại hình đơn lập có 6 thanh điệu. Vì vậy, khác với âm tiết các ngôn ngữ châu Âu, âm tiết nào của tiếng Việt cũng mang một thanh điệu nhất định. Trong dòng lời nói, âm tiết tiếng Việt bao giờ cũng thể hiện khá đầy đủ, rõ ràng, được tách và ngắt ra thành từng khúc đoạn riêng biệt.

Theo Cao Xuân Hạo trong [12], hệ thống âm vị tiếng Việt bao gồm 22 phụ âm đầu, 6 phụ âm cuối, 16 nguyên âm và 2 bán nguyên âm. Chi tiết hệ thống âm vị như sau:

- 22 phụ âm đầu: **/b, m, f, v, t, t', d, n, z, ʐ, s, ʂ, c, ʈ, ɲ, l, k, χ, ŋ, ɣ, h, ʔ**/
- 6 phụ âm cuối : **/m, n, ŋ, p, t, k/**
- 2 bán nguyên âm: **/-w, -j/**
- 16 nguyên âm: **/i, e, ɛ, ɤ, ɤ̆, a, ɯ, ă, u, o, ɔ, ɔ̆, ɛ̆, ie, ɯɤ, uo/**

Bảng 2.3 Bảng phụ âm đầu

Phương thức \ Vị trí				Môi	Đầu lưỡi		Mặt lưỡi	Gốc lưỡi	Thanh hầu
					Bọt	Lưỡi			
Tắc	Ồn	Bật hơi			t'				
		Không bật hơi	Vô thanh		t	ʈ	c	k	ʔ
			Hữu thanh	b	d				
	Vang			m	n		ɲ	ŋ	
Xát	Ồn	Vô thanh		f	s	ʂ		χ	h
		Hữu thanh		v	z	ʐ		ɣ	
	Vang				l				

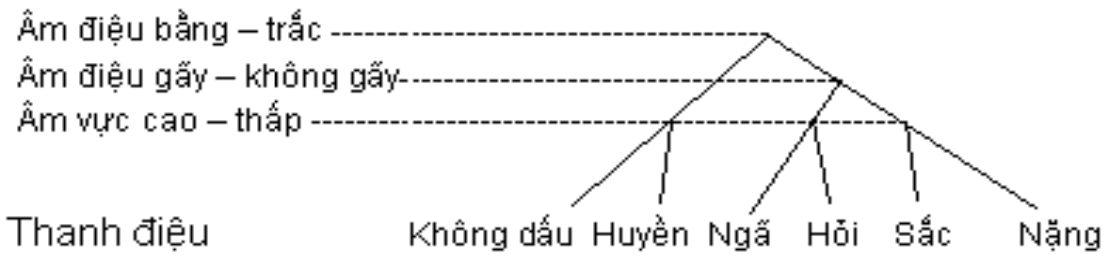
Bảng 2.4 Bảng phụ âm cuối và bán nguyên âm

Phương thức \ Vị trí		Môi	Lưỡi	
			Đầu lưỡi	Gốc lưỡi
Ồn		p	t	k
Vang	Mũi	m	n	ŋ
	Không mũi	-w	-j	

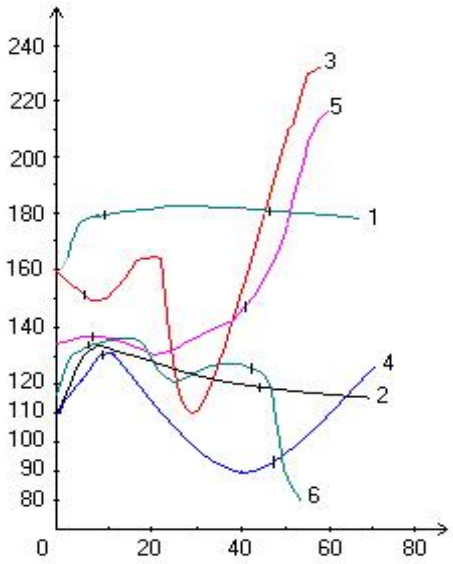
Bảng 2.5 Bảng nguyên âm

Âm sắc	Vị trí lưỡi, hình dáng môi / Độ mở của miệng	Trước, không tròn môi	Sau	
			Không tròn môi	Tròn môi
Cố định	Nhỏ	i	ɯ	u
	Lớn vừa	e	ɤ/ɤ̆	o
	Lớn	ɛ/ɛ̆	a/ă	ɔ/ɔ̆
Không cố định		iẹ	ɯɤ	uọ

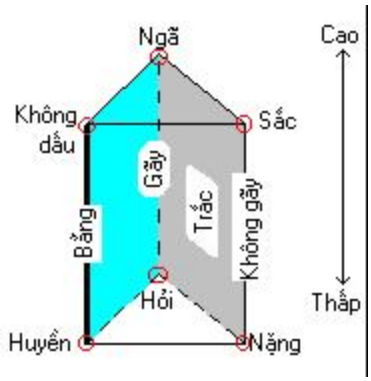
Theo Mai Ngọc Chữ [7, tr 91-105] tiếng Việt có 6 thanh điệu gồm: không dấu, huyền, ngã, hỏi, sắc, nặng.



Hình 2.2 Sơ đồ về 3 tiêu chí khu biệt cho sáu âm vị thanh điệu



Hình 2.3 Hình minh họa biểu đồ thanh điệu



Hình 2.4 Hình minh họa lăng trụ thanh điệu

2.1.2.2. Hình vị

Hình vị là một hoặc chuỗi kết hợp một vài âm vị, biểu thị một khái niệm. Hình vị là đơn vị nhỏ nhất có ý nghĩa. Chức năng của hình vị là chức năng ngữ nghĩa.

Ví dụ, kết hợp "quốc gia" trong tiếng Việt gồm hai hình vị: "quốc" là nước, "gia" là nhà; "паровоз" của tiếng Nga gồm ba hình vị "пар" là hơi nước, "воз" là sự chuyên chở, còn "-о" là hình vị nối.

2.1.2.3. Từ

Từ là một khái niệm đã từng được rất nhiều nhà ngôn ngữ học định nghĩa. Từ thời Hy Lạp cổ đại, trường phái ngôn ngữ Alexandri đã định nghĩa: “Từ là đơn vị nhỏ nhất trong chuỗi lời nói”.

Theo E.Sapir thì: “Từ là một đoạn nhỏ nhất có ý nghĩa, hoàn toàn có khả năng độc lập và bản thân có thể làm thành câu tối giản”.

Theo L.Bloomfield thì từ là “một hình thái tự do nhất”.

Theo B.Golovin thì từ là “đơn vị nhỏ nhất có ý nghĩa của ngôn ngữ, được vận dụng độc lập, tái hiện tự do trong lời nói để xây dựng nên câu”. Đây cũng chính là định nghĩa mà trong ngôn ngữ học đại cương hay sử dụng.

Theo Solnev thì “Từ là đơn vị ngôn ngữ có tính hai mặt: âm và nghĩa. Từ có khả năng độc lập về cú pháp khi sử dụng trong lời”.

Theo Lục Chí Vĩ thì "Từ là đơn vị nhỏ nhất có thể vận dụng tự do trong câu".

Theo V.G.Admoni thì "Từ là đơn vị ngữ pháp, do hình vị cấu tạo nên, dùng để biểu thị đối tượng, quá trình, tính chất và những mối quan hệ trong hiện thực, có tính đặc thù rõ rệt và có khả năng kiến lập nhiều mối quan hệ đa dạng nhau".

Theo R.A.Bundagop thì "Từ là đơn vị nhỏ nhất và độc lập, có hình thức vật chất và có nghĩa, có tính chất biện chứng và lịch sử".

Trong ngôn ngữ tiếng Việt, cũng có nhiều định nghĩa được đưa ra. Theo quan điểm của Trương Văn Trình và Nguyễn Hiến Lê thì "Từ là âm có nghĩa, dùng trong ngôn ngữ để diễn tả một ý đơn giản nhất, nghĩa là ý không thể phân tích ra được".

Theo Phan Khôi thì "Từ là một lời để tỏ ra một khái niệm trong khi nói".. Theo Nguyễn Lân thì "Từ là những tiếng có nghĩa, tức là mỗi khi nghe thấy, trong óc chúng ta đều có một khái niệm".

Theo Nguyễn Kim Thản thì "Từ là đơn vị cơ bản của ngôn ngữ, có thể tách khỏi các đơn vị khác của lời nói để vận dụng một cách độc lập và là một khối hoàn chỉnh về mặt ý nghĩa (từ vựng hay ngữ pháp) và cấu tạo". Quan niệm cơ bản của ông về "đơn vị cơ bản" là những đơn vị có số lượng hữu hạn và có nghĩa. Đơn vị đó không thể là câu (vì số lượng câu là vô hạn) và cũng không thể là âm tiết (vì nhiều âm tiết không có nghĩa). Vậy đơn vị cơ bản là cái gì đó nhỏ hơn câu và lớn hơn âm tiết.

Theo Hồ Lê thì "Từ là đơn vị ngôn ngữ có chức năng định danh phi liên kết hiện thực, hoặc có chức năng mô phỏng tiếng động, có khả năng kết hợp tự do, có tính vững chắc về cấu tạo và tính nhất thể về ý nghĩa". Theo ông từ khác với âm tiết chủ yếu về mặt ý nghĩa. Từ có khả năng kết hợp tự do khi sử dụng. Từ khác cụm từ bởi tính vững chắc về cấu tạo, tính nhất thể về ý nghĩa.

Nguyễn Tài Cẩn tuy không trực tiếp định nghĩa từ trong tiếng Việt, nhưng ông đã chứng minh những tính chất đặc biệt của "tiếng", một đơn vị được ông xem như là hình vị và có tính năng rất gần với từ, nó cũng có thể là "từ đơn" và là thành tố trực tiếp tạo nên "từ ghép". Theo ông, mọi đặc thù về từ của tiếng Việt bắt nguồn từ

tính đơn lập của tiếng Việt mà thể hiện rõ nét nhất là qua một đơn vị đặc biệt, đó là "tiếng". Quan điểm này cũng được Cao Xuân Hạo đồng tình.

Kế thừa quan điểm coi tiếng gần là từ, Nguyễn Thiện Giáp đã phát triển tư tưởng này, ông coi "tiếng" trong tiếng Việt chính là từ trong các ngôn ngữ Ấn-Âu.

Theo Mai Ngọc Chữ "Từ là đơn vị nhỏ nhất có nghĩa, có kết cấu vỏ ngữ âm bền vững, hoàn chỉnh, có chức năng gọi tên, được vận dụng độc lập, tái hiện tự do trong lời nói để tạo câu".[10]

Có thể thấy, có rất nhiều quan điểm về khái niệm "từ", những quan điểm này tuy có những khác biệt, nhưng hầu như không đối lập mà bổ sung cho nhau. Vậy chính xác từ là gì? Viện sĩ L.V.Sherba đã phát biểu: "Trong thực tế, từ là gì? Thiết nghĩ rằng trong các ngôn ngữ khác nhau, từ sẽ khác nhau. Do đó tất yếu sẽ không có khái niệm từ nói chung". [23]Cho đến nay vẫn chưa có một định nghĩa nào trọn vẹn về từ, Đinh Điền và Hồ Bảo Quốc trong [10] đã chỉ ra những nét đặc trưng chính của từ như sau:

- Về hình thức: từphải là một khối vềcấu tạo (mặt chính tả, mặt ngữâm,…)
- Về nội dung: từphải có ý nghĩa hoàn chỉnh.
- Về khả năng: từcó khảnăng hoạt động tựdo và độc lập vềcú pháp.

Chúng tôi thừa nhận và sử dụng những nét đặc trưng trên làm tiêu chí nhận diện từ trong luận văn này.

2.1.2.4. Câu

Câu là chuỗi kết hợp của một hay nhiều từ, chức năng của nó là chức năng thông báo.

2.1.3 Cấu trúc của đơn vị từ tiếng Việt

2.1.3.1 Tiếng

"Tiếng" (còn gọi là âm tiết) là đơn vị cơ bản trong tiếng Việt dùng để cấu tạo các đơn vị ngôn ngữ khác cao hơn. Số lượng tiếng trong tiếng Việt không lớn (khoảng 10.000), và chiều dài mỗi tiếng ngắn (không quá 7 chữ cái). Trong xử lý

tiếng Việt tự động bằng máy tính, thì “tiếng” là đơn vị tự nhiên nhất mà máy tính dễ dàng lưu trữ, nhận diện và xử lý. Tiếng chính là “từ chính tả”. [10]

Từ tiếng Việt được cấu tạo bằng các dùng một tiếng hoặc tổ hợp các tiếng. Khi dùng một tiếng tạo nên từ ta được các từ đơn. Khi tổ hợp các tiếng để tạo nên từ ta được các từ phức. Ngoài ra còn có cụm từ cố định được cấu tạo bởi sự kết hợp chặt chẽ các từ khác.

2.1.3.2Từ đơn

Từ đơn là từ mà bộ phận không thể chia nhỏ, đủ nghĩa, có thể đứng độc lập trong câu. Nói cách khác, từ đơn là từ chỉ có một thành tố, mỗi thành tố là một hình vị.

Căn cứ vào số lượng âm tiết có thể chia từ đơn ra làm hai loại: từ đơn đơn âm và từ đơn đa âm.

- Từ đơn đơn âm: gồm một hình vị, mỗi hình vị là một âm tiết. Ví dụ: trời, đất, nhà, quán, vui, v.v…
- Từ đơn đa âm: gồm một hình vị, hình vị này gồm 2 âm tiết trở lên, ví dụ: cà vạt, cà phê, tivi, …

2.1.3.3 Từ phức

Từ phức là từ được cấu tạo bởi hai hay nhiều hình vị kết hợp. Ví dụ: hải quân, bộ binh, nhà nước, Thảo Cầm Viên, v.v… Từ phức được chia làm 2 loại: từ láy và từ ghép.

2.1.3.3.1 Từ láy

Từ láy là những từ có một hình vị gốc và một hình vị láy. Hình vị láy có dạng ngữ âm trùng lắp hoàn toàn hoặc bộ phận với hình vị gốc. Có thể phân chia từ láy dựa vào số lượng âm tiết thành các loại như sau:

- Láy đôi: đo đỏ, chầm chậm, lung linh, …
- Láy ba: sạch sành sanh, tuốt tuồn tuột, …
- Láy tư: bối hối bồi hồi, khập khà khập khiễng, …

Ngoài ra từ láy còn có thể chia theo láy toàn phần và láy bộ phận:

- Láy toàn phần: xanh xanh, đỏ đỏ, vui vui, buồn buồn, ầm ầm, …
- Láy bộ phận: gồm láy âm và láy vần
    - Láy âm: phụ âm đầu của hình vị gốc được lập lại trong hình vị láy, còn vần thì thay đổi. Ví dụ: dễ dàng, gòn gàng, chim chóc, mạnh mẽ, …
    - Láy vần: phần vần của hình vị gốc được lập lại trong hình vị láy, phần phụ âm đầu thì thay đổi. Ví dụ: lúng túng, bối rối, bèo nhèo, …

2.1.3.3.2 Từ ghép

❖ **Khái niệm:**

Từ ghép là từ có ít nhất hai thành tố kết hợp với nhau. Mỗi thành tố có thể là một hình vị đơn âm, đa âm hay thậm chí là một tổ hợp hình vị. Đồng thời mỗi thành tố đều phải có nghĩa.

Ví dụ: xe đạp: có 2 thành tố là [xe, đạp]. Cà phê chè: gồm 2 thành tố là [cà phê, chè]. Trong đó cà phê: hình vị đa âm tiết, chè: hình vị đơn âm tiết.

Có những trường hợp hình thức có vẻ giống từ láy, nhưng thực ra là từ ghép vì các thành tố cấu thành đều có nghĩa riêng. Ví dụ: bóng bay, song sắt, v.v...

❖ **Tính chất cơ bản của từ ghép:**

Sự kết hợp các hình vị tạo thành từ ghép phải chặt chẽ về mặt hình thức và cả về mặt ngữ nghĩa.

- Chặt chẽ về hình thức: thể hiện ở chỗ trong tiếng Việt các hình vị trong từ ghép luôn đi với nhau, từng hình vị không thể tách rời nhau. Nếu không thỏa mãn yêu cầu thì ý nghĩa ban đầu của từ ghép bị phá vỡ.

  Ví dụ: các từ nhà gỗ, bàn cây, ghế sắt chúng ta có thể diễn đạt lại như sau: cái nhà làm bằng gỗ, cái bàn làm bằng cây, cái ghế làm bằng sắt. Có thể bỏ đi các từ “làm bằng” thì nghĩa vẫn không thay đổi. Nhưng từ nhà gỗ nếu bỏ đi 1 trong 2 từ [nhà, gỗ] thì ý nghĩa sẽ khác hẳn.

– Chặt chẽ về ngữ nghĩa: có nhiều mức độ khác nhau, mức độ cao nhất là có tính thành ngữ. Tính thành ngữ được hiểu khi ý nghĩa của một tổ hợp không thể giải thích bằng cách giải nghĩa của từng yếu tố tạo nên nó.

Ví dụ: khi nói về một thầy thuốc có tài chữa bệnh, ta có thể dùng từ: mát tay. Nhưng khi tách các từ ra để giải nghĩa thì ta lại thu được ý nghĩa khác với ban đầu.Một số ví dụ khác như: lắm miệng (nhiều chuyện), yếu tim (nhát), …

2.1.3.4Cụm từ cố định

Cụm từ cố định có thể chia thành 3 loại: thành ngữ, quán ngữ và ngữ cố định định danh. [8],[17]

2.1.3.4.1 Thành ngữ

Thành ngữ là cụm từ cố định, hoàn chỉnh về cấu trúc và ý nghĩa. Nghĩa của chúng có tính hình tượng hoặc/và gợi cảm.Ví dụ: ba cọc ba đồng, chó cắn áo rách, nhà ngói cây mít, bán bò tậu ễnh ương, méo miệng đòi ăn xôi vò, ông mất của kia bà chìa của nọ, đủng đỉnh như chĩnh trôi sông, v.v...

2.1.3.4.2Quán ngữ

Quán ngữ là những cụm từ được dùng lặp đi lặp lại trong các loại diễn từthuộc phong cách khác nhau. Chức năng của chúng là để đưa đẩy, rào đón, để nhấn mạnh hoặc để liên kết trong diễn từ.Ví dụ: của đáng tội, bỏ ngoài tai, nói tóm lại, kết cục là, nói cách khác,v.v...

2.1.3.4.3. Ngữ cố định định danh

Ngữ cố định định danh là những cụm từ cố định, định danh, gọi tên sự vật, là những đơn vị ổn định về cấu trúc và ý nghĩa hơn các quán ngữ, nhưng ý nghĩa mang tính hình tượng chưa được như thành ngữ.Trong mỗi cụm từ như vậy thường có một thành tố chính và một vài thành tố phụ miêu tả sự vật được nêu ở thành tố chính.

Ví dụ: lông mày lá liễu, mắt bồ câu, trẻ măng, tóc rễ tre, con gái rượu, bàn mưu tính kế, v.v...

2.1.4 Từ vựng tiếng Việt

2.1.4.1 Giới thiệu

Từ vựng là tập hợp tất cả các từ và những đơn vị tương đương với từ của một ngôn ngữ. Từ vựng là chất liệu cần thiết, cơ bản nhất để kiến tạo nên một ngôn ngữ mà nếu thiếu nó chúng ta không thể hình dung được ngôn ngữ này [4].

2.1.4.2 Sự phân lớp từ vựng

Hệ thống từ vựng của một ngôn ngữ thường rất lớn, để thuận tiện cho việc học tập và nghiên cứu người ta có thể phân chia thành các lớp từ vựng riêng mang ý nghĩa khái quát chung của một tập hợp, một nhóm. Theo nghiên cứu của Mai Ngọc Chừ và cộng sự trong [8] trong ngôn ngữ tiếng Việt các lớp từ vựng được phân lớp dựa vào các tiêu chí sau:

- Theo nguồn gốc
- Theo phạm vi sử dụng
- Theo tần số sử dụng
- Theo phong cách sử dụng

2.1.4.2.1 Theo tiêu chí nguồn gốc

Theo tiêu chí nguồn gốc, từ vựng thường được chia làm hai lớp: lớp từ thuần và lớp từ ngoại lai. Lớp từ thuần là lớp từ vốn có của ngôn ngữ đó; còn lớp từ ngoại lai là lớp từ vay mượn của ngôn ngữ khác trong quá trình giao thao văn hóa.

Trong tiếng Việt, có lớp từ thuần Việt và lớp từ có nguồn gốc từ tiếng Hán (gồm Hán Việt và Hán cổ), gốc Ấn - Âu (tiếng Anh, tiếng Pháp, tiếng Nga v.v…). Có thể liệt kê một số từ thông dụng như: khăn mùi xoa, xà phòng, sô cô la, ti vi, mít tin, căn tin, cà vạt, bi đông, …

2.1.4.2.2 Theo tiêu chí phạm vi sử dụng

Theo tiêu chí phạm vi sử dụng, từ vựng tiếng Việt được chia thành các lớp: từ phổ thông, từ địa phương, từ nghề nghiệp, thuật ngữ, tiếng lóng.

Từ phổ thông: là lớp từ được đại đa số mọi người trong cộng đồng sử dụng. Mỗi ngôn ngữ đều có lớp từ này, đóng vai trò cơ bản trong hệ thống từ vựng của một ngôn ngữ. Lớp từ này chính là lớp từ chuẩn thường sử dụng trong văn viết của ngôn ngữ đó.

Từ địa phương: là lớp từ thuộc một phương ngữ, thường dùng trong giao tiếp hàng ngày, chỉ phổ biến trong lãnh thổ, phạm vi địa phương đó. Ví dụ: thầy, u, mền, tía, má, …

Thuật ngữ: là những từ ngữ làm tên gọi cho các khái niệm, các đối tượng xác định trong mỗi ngành, mỗi lĩnh vực khoa học. Có tính chất: chính xác, chuẩn tắc, hệ thống và quốc tế hóa. Ví dụ một số thuật ngữ trong lĩnh vực hóa học: đơn chất, hợp chất, hữu cơ, nguyên tử, v.v...

Từ nghề nghiệp: là lớp từ bao gồm những đơn vị từ ngữ được sử dụng phổ biến trong phạm vi ngành nghề nào đó. Ví dụ ở lĩnh vực nghề làm mộc có những từ như: bào cóc, bào xoa, mộng, v.v... Ở lĩnh vực hát tuồng có: đào, kép, v.v...

Tiếng lóng: là lớp từ do những nhóm người trong xã hội dùng để gọi tên những sự vật, hiện tượng, hành động vốn đã có tên gọi trong vốn từ vựng chung. Ví dụ: lính phòng không (trai chưa vợ), hôi của (lấy đồ của người khác khi xảy ra sự cố), xế hộp (xe ô tô đắt tiền) v.v...

2.1.4.2.3 Theo tiêu chí tần số sử dụng

Theo tiêu chí tần số sử dụng, từ vựng tiếng Việt được phân thành hai lớp: từ ngữ tích cực và từ ngữ tiêu cực.

Từ ngữ tích cực: là những từ ngữ được mọi người sử dụng ở mọi lúc, mọi nơi, có tần số xuất hiện cao, độ phân bố lớn. Đây là thành phần cơ bản của từ vựng.

Từ ngữ tiêu cực: là những từ có tần số sử dụng thấp. Chia làm hai loại:

+ Từ mới: là những từ xuất hiện để bù đắp sự thiếu hụt của từ vựng. Khi mới xuất hiện, từ mới thường không được sử dụng rộng rãi nên thuộc lớp từ ngữ tiêu cực. Khi từ mới được chấp nhận và được sử dụng phổ biến thì được trở thành từ

ngữ tích cực. Ví dụ một số từ ngữ trước đây thuộc lớp tiêu cực: tổ chức (làm tiệc), xây dựng (lập gia đình), phần mềm, phần cứng, v.v...

+ Từ cũ: là những từ bị loại dần khỏi hệ thống từ vựng hiện tại bởi các nguyên nhân lịch sử, xã hội, văn hóa, v.v... Ví dụ: điền trang (trang trại lớn), thái thú (một chức quan), dân cày (người làm ruộng), gác bờ đu (cái chắn bùn), v.v...

2.1.4.2.4 Theo tiêu chí phong cách sử dụng

Theo tiêu chí phong cách sử dụng, từ vựng tiếng Việt được phân thành ba lớp: lớp từ khẩu ngữ, lớp từ thuộc phong cách viết và lớp từ trung hòa.

Lớp từ khẩu ngữ: là những từ ngữ dùng trong giao tiếp, thường có những đặc điểm sau đây: tự do, phóng túng, cường điệu, thường dùng kèm thành ngữ, quán ngữ, các từ thưa gửi, v.v... Ví dụ: lo thắt ruột, chờ đỏ mắt, đánh sặc tiết, chạy bở hơi tai, v.v...

Lớp từ thuộc phong cách viết: là những từ được chọn lọc, trau dồi, gắn bó với chuẩn tắc nghiêm ngặt. Có đặc điểm chung là gắn liền với nội dung của một số phong cách chức năng cụ thể như: phong cách khoa học, hành chính sự vụ, chính luận báo chí, văn học. Không mang tính thông tục. Mang tính khái quát, trừu tượng, v.v... tuỳ theo phạm vi riêng của mỗi phong cách chức năng. Thường dùng nhiều các từ có gốc Hán, Ấn-Âu được du nhập. Ví dụ: phong cách khoa học: âm vị, hình vị, ngữ pháp, v.v... ; phong cách hành chính sự vụ: công văn, văn thư, tố tụng, v.v... ; phong cách văn học: đắm đuối, lộng lẫy, v.v...

Lớp từ trung hòa: là những từ không mang dấu hiệu đặc trưng của lớp từ khẩu ngữ hoặc lớp từ thuộc phong cách viết. Ví dụ: đau buồn, lặng lẽ, đi tản bộ, v.v...

2.1.5 Vấn đề nhập nhằng nghĩa của từ

2.1.5.1 Giới thiệu

Nhập nhằng là hiện tượng mơ hồ về ngữ nghĩa, không phân định rạch ròi ranh giới giữa các từ do hiện tượng đa nghĩa, đa từ loại của từ, hoặc do sự kết hợp của các âm tiết đặt cạnh nhau tạo thành những từ khác nhau, v.v… Đây là hiện tượng

thường gặp khi xử lý ngôn ngữ tự nhiên. Để xử lý nhập nhằng đòi hỏi người xử lý cần hiểu đúng nghĩa của từ trong câu. Với con người thì đây không phải là vấn lớn, vì con người có thể hiểu đúng nghĩa của từ trong câu nhờ vào các yếu tố khác như: ngữ cảnh, người nói, người nghe, hoàn cảnh lịch sử, v.v… Nhưng với máy tính thì đây lại là một vấn đề nghiêm trọng, vì máy tính không phải là con người.

Ví dụ cho câu: Con bò cạp con bò cạp.

Câu này có thể hiểu theo nhiều cách:

- Con bò cạp | con bò cạp

- Con bò | cạp | con bò cạp

- Con bò | cạp | con bò | cạp

2.1.5.2 Một số hiện tượng nhập nhằng

2.1.5.2.1 Nhập nhằng ranh giới từ

Về ranh giới từ, sự nhập nhằng của tiếng Việt có thể chia thành 2 kiểu sau:

- Nhập nhằng chồng chéo: chuỗi “abc” được gọi là nhặp nhằng chồng chéo nếu như từ“ab”, “bc” đều xuất hiện trong từ điển tiếng Việt.

Ví dụ như trong câu “ông già đi nhanh quá” thì chuỗi “ông già đi” bị nhập nhằng chồng chéo vì các từ “ông già” và “già đi” đều có trong từ điển.

- Nhập nhằng kết hợp: chuỗi “abc” được gọi là nhập nhằng kết hợp nếu như từ “a”, “b”, “ab” đều xuất hiện trong từ điển tiếng Việt.

Ví dụ như trong câu: “Bàn là này còn rất mới” thì chuỗi “bàn là” bị nhập nhằng kết hợp, do các từ “bàn”, “là”, “bàn là” đều có trong từ điển.

2.1.5.2.2 Nhập nhằng từ đa nghĩa

Bất cứ ngôn ngữ nào cũng có từ đa nghĩa, nguyên nhân là vì rất nhiều khái niệm có các sắc thái ý nghĩa tuy không hoàn toàn trùng khớp nhau nhưng lại có nhiều nét tương đồng.

Ví dụ: Cho 2 câu

- Câu 1: Nó đã ăn bánh ở đây

- Câu 2: Nó đã ăn cướp ở đây.

Có thể thấy chữ "ăn" ở câu 1 và chữ "ăn" ở câu 2 có ý nghĩa hoàn toàn khác nhau. Ngoài ra cách dùng cũng khác nhau. Ở câu 1, chữ "ăn" là một từ, chỉ hành động ăn uống. Trong khi ở câu 2, chữ "ăn" lại là một âm tiết trong từ "ăn cướp", chi hành động xấu.

Hiện tượng này gây cản trở khá lớntrong xử lý ngôn ngữ tự nhiên như phân đoạn từ, dịch tự động, ...

2.1.5.2.3 Nhập nhằng từ đồng âm

Hai từ đồng âm với nhau nghĩa là hai từ có âm giống nhau nhưng mang nghĩa khác nhau, còn đồng tự là hai từ về mặt ký tự là giống nhau nhưng nghĩa khác nhau. Do đặc điểm của tiếng Việt từ đồng âm cũng thường là từ đồng tự, ở các ngôn ngữ khác hai hiện tượng này không trùng khớp nhau. Cũng phải phân biệt từ đồng tự với từ đa nghĩa, trong từ đa nghĩa các nghĩa đều có chung một nguồn gốc và do vậy luôn có nét tương đồng trong khi đó trong từ đồng tự chúng không có liên hệ về nguồn gốc với nhau, nghĩa của chúng khác nhau rõ rệt. Ví dụ từ kiếm trong hai câu sau đây là hai từ đồng tự:

- **Anh ta sử dụng kiếm rất điêu luyện.**
- **Kiếm ăn bây giờ khó lắm.**

Việc xác định nghĩa chính xác của từ đồng tự dễ dàng hơn từ đa nghĩa bởi vì sự khác nhau lớn về ngữ nghĩa của chúng giúp đưa ra được nhiều tiêu chuẩn tốt để phân biệt.

2.1.5.2.4 Nhập nhằng từ loại

Từ loại là một yếu tố quan trọng trong việc xác định nghĩa chính xác của từ và sắp xếp các từ thành câu hoàn chỉnh trong dịch tự động. Như vậy có nghĩa là từ loại giúp khử nhập nhằng, nhưng chính bản thân nó trong một số trường hợp cũng nhập nhằng.

Phần lớn các ngôn ngữ biến hình từ loại được xác định tương đối dễ dàng vì khi chuyển loại thì từ cũng chuyển kiểu hình của nó ví dụ trong tiếng Anh từ free là tính từ có nghĩa là tự do, chuyển loại thành danh từ có thêm hậu tố "dom" thành freedom nghĩa là sự tự do. Điều này tạo thuận lợi cho việc gán nhãn từ loại một cách tự động nhờ các dấu hiệu nhận biết tổng quát.

Các ngôn ngữ không biến hình như tiếng Việt vấn đề xác định từ loại yêu cầu các thuật toán phức tạp hơn,bắt buộc phải phân tích cú pháp, mặt khác ngay trong nội bộ ngành ngôn ngữ vẫn chưa có sự thống nhất về phân loại từ loại cho tiếng Việt.

2.2 Cơ sở lý thuyết về ngôn ngữ học thống kê

2.2.1 Tổng quan về ngôn ngữ học thống kê

Ngôn ngữ học thống kê là một ngành khoa học có truyền thống lâu đời, ra đời trên cơ sở nghiên cứu về ngôn ngữ kết hợp với lý thuyết xác suất thống kê. Ngay từ thế kỷ 18, F.Kaeding đã áp dụng phương pháp thống kê trong ngôn ngữ để xây dựng từ điển tần số tiếng Đức. Năm 1913, nhà toán học Nga A.A.Markov đã dùng phương pháp xác suất thống kê để nghiên cứu về quy luật nối tiếp nhau của các phụ âm và nguyên âm trong tiếng Nga.

Từ thập niên 1950 trở lại đây, ngôn ngữ học thống kê đã liên tục phát triển và đạt được nhiều thành tựu ở khắp các lĩnh vực ngôn ngữ học như: ngữ âm học, từ vựng học, ngữ pháp học, ngữ nghĩa học, v.v... và đặc biệt là trong lĩnh vực máy học (ML: Machine Learning).

Ngôn ngữ học thống kê đã được áp dụng trong nhiều bài toán ngôn ngữ như:

- Nghiên cứu loại hình của ngôn ngữ.
- Xây dựng từ điển từ vựng, từ điển tần số của một ngôn ngữ.
- Xác định văn phong của tác giả thông qua các tác phẩm.
- Xử lý thông tin tự động như: tách câu, tách từ, dịch máy, sửa lỗi chính tả, nhận dạng tiếng nói.
- Xác định niên đại của ngôn ngữ, v.v...

Cơ sở toán học của ngôn ngữ học thống kê chính là lý thuyết xác suất thống kê. Trong phần tiếp theo chúng tôi sẽ trình bày một số lý thuyết thống kê thường dùng trong xử lý ngôn ngữ.

2.2.2Một số lý thuyết xác suất thống kê trong xử lý ngôn ngữ

2.2.2.1 Hàm xác suất

Hàm xác suất P của một biến ngẫu nhiên E là một ánh xạ từ miền xác định của E (không gian các giá trị E có thể nhận) đến đoạn số thực [0,1].

Giả sử E có thể nhận các giá trị phân biệt e1, e2, ... , en.

Hàm xác xuất phải thỏa các tính chất sau:

1. $P(e_i) \geq 0, \forall i$
2. $P(e_i) \leq 1, \forall i$
3. $\sum_{i=1}^{n} P(e_i) = 1$

2.2.2.2 Xác suất điều kiện

Cho các biến ngẫu nhiên X và Y, xác xuất điều kiện được định nghĩa:

$$P(X|Y) = \frac{P(XY)}{P(Y)}$$

Trong đó sử dụng ký hiệu:

1. P(X) thay cho P(X=x).
2. P(XY) có nghĩa là đồng thời có X=x và Y=y.

❖ **Xác suất đồng thời**

$$P(XY) = P(X) * P\left(\frac{Y}{X}\right)$$

❖ **Định luật Bayes**

$$P\left(\frac{X}{Y}\right) = \frac{P\left(\frac{Y}{X}\right) * P(X)}{P(Y)}$$

❖ **Biến ngẫu nhiên độc lập**

Hai biến ngẫu nhiên X, Y được gọi là độc lập khi và chỉ khi:

—

Từ đó suy ra nếu X, Y độc lập thì:

❖ **Kỳ vọng và phương sai**

Kỳ vọng là giá trị trung bình của biến ngẫu nhiên. Giả sử X là biến ngẫu nhiên, thì kỳ vọng là:

$$E(X) = \sum_x xP(x)$$

Phương sai của biến ngẫu nhiên A là một số không âm dùng để đo mức độ phân tán các giá trị của biến ngẫu nhiên xung quanh giá trị trung bình của nó.

$$Var(X) = E\left(\left(X - E(X)\right)^2\right) = E(X^2 - E^2(X))$$

2.2.2.3 Xác suất ước lượng

❖ **Xác suất chính xác**

Trong lý thuyết xác suất, nếu có đầy đủ dữ liệu ta có thể tính toán xác suất chính xác của một biến ngẫu nhiên. Chẳng hạn với đầy đủ số liệu thống kê từ đầu năm (01-01-2014) đến ngày hôm nay (31-05-2014) về sự kiện "trời mưa trong ngày", đếm được 39 ngày mưa trên tổng số 150 ngày. Như vậy ta tính được xác suất chính xác xảy ra sự kiện "trời mưa trong ngày" là 39/150 trong khoảng thời gian từ đầu năm (01-01-2014) đến hết ngày hôm nay (31-05-2014).

Tuy nhiên, xác suất chính xác không phải là ứng dụng chính của lý thuyết xác suất thống kê. Vì ta chưa thể tính được xác suất "trời mưa trong ngày" cho những ngày kế tiếp, tại những thời điểm chưa thu thập được thông tin chính xác.

❖ **Ước lượng khả năng cực đại (MLE: Maximum Likelihood Estimator)**

Là cách ước lượng dùng xác suất chính xác để ước lượng cho một biến cố chưa biết. Cách ước lượng MLE có độ chính xác tùy thuộc vào độ lớn của dữ liệu: càng lớn càng chính xác.

❖ **Ước lượng khả năng mong muốn (ELE: Expect Likelihood Estimator)**

Đây là khả năng ước lượng thích hợp cho loại biến cố có tần suất thấp. Xét biến ngẫu nhiên X, gọi Vi là số lần xuất hiện X = xi.

Kỹ thuật ước lượng ELE tính xác suất theo công thức sau:

$$P(X = X_i) \cong (\frac{V_i}{\Sigma i V_i})$$

Để tránh vấn đề tử số bằng không, (xác suất bằng không), ta cộng thêm một gia số Vi, chẳng hạn:

$V_i = |X_i| + 0.5$, trong đó $|X_i|$ là số lần $X = x_i$

Chúng ta xét tình huống dưới đây để phân biệt MLE và ELE:

Giả sử trong tiếng t không xuất hiện trong kho ngữ liệu và ta cần ước lượng xác suất suất hiện của t trong các từ w1, w2, ... w40.

Như vậy ta có một biến ngẫu nhiên X, với X = xi chỉ khi t xuất hiện trong từ wi.

Với công thức ước lượng MLE, ta có P(X = xi) không xác định (=0), nghĩa là kho ngữ liệu không cung cấp thông tin gì về sự xuất hiện của tiếng t trong từ wi.

Ngược lại, với công thức ELE, ta có $V_i = 0.5, \forall i = 1..40$, do đó:

$$P\left(\frac{W_i}{t}\right) \cong \frac{0.5}{0.5 * 40} = 0.025$$

Công thức này thể hiện thông tin ước lượng sự xuất hiện của t trong wi, mặc dù trong kho ngữ liệu không hề có *t.*

2.2.2.4Mô hình Markov ẩn và các mô hìnhN-Gram

❖ **Mô hình Markov ẩn**

Mô hình Markov ẩn (HMM: Hidden Markov Model)[1][26] là một tiến trình ngẫu nhiên kép. Tầng ngẫu nhiên đầu tiên là tiến trình Markov bậc một được biểu diễn bởi một lược đồ chuyển trạng thái . Mỗi trạng thái là một quan sát có thể của tiến trình Makov và xác suất chuyển từ trạng thái A sang thạng thái B là $P(St + 1 = B \mid St = A)$ – xác suất đến trạng thái B ở thời điểm $t + 1$ với điều kiện ở thời điểm t ở thạng thái A. Tầng ngẫu nhiên thứ hai là tập các xác suất nhả của mỗi trạng thái. Tầng xác suất thứ hai này tạo ra một bức màn chắn, tức là với một dãy ký hiệu quan sát được, dãy trạng thái sinh ra dãy ký hiệu đó mơ hồ (không quan sát được). Nó được ẩn đối với người quan sát.

HMM đã và đang được sử dụng phổ biến cho các mô hình thống kê, đặc biệt là trong lĩnh vực xử lý ngôn ngữ tự nhiên. HMM thực chất là một hàm xác suất của quá trình Markov. Các quá trình Markov được phát triển đầu tiên bởi Andrei A.Markov, lần đầu tiên được ứng dụng cho mục đích xử lý ngôn ngữ – mô hình các chuỗi ký tự trong tác phẩm văn học Nga (Markov 1913) – nhưng sau đó mô hình này đã được phát triển thành một công cụ thống kê tổng quát.

Các mô hình Markov có thể sử dụng khi cần mô hình hóa xác suất của một dãy tuyến tính các sự kiện . Chẳng hạn, chúng được sử dụng trong xử lý ngôn ngữ tự nhiên để mô hình hóa dãy các ký tự (xử lý văn bản), âm tiết (xử lý tiếng nói), v.v...

Trong một HMM, ta không biết dãy trạng thái mà mô hình đi qua, nhưng biết một hàm xác suất của nó. Sau đây là dạng thức tổng quát của một HMM:

HMM là một bộ 5 (S, K, π, A, B)

Trong đó:

1. S: tập các trạng thái
2. K: tập các ký tự output
3. π: xác suất đối với trạng thái ban đầu
4. A: xác suất sự dịch chuyển trạng thái

5. B: xác suất phát ra ký hiệu

Bảng 2.6 Bảng liệt kê các ký hiệu thường được sử dụng trong HMM

Tập các trạng thái	$S = \{1, \ldots, N\}$
Mẫu ký tự output	$K = \{1, \ldots, M\} = \{1, \ldots, M\}$
Các xác suất trạng thái ban đầu	$\pi = \{ \pi_i \}, i \in S$
Các xác suất chuyển dịch trạng thái	$A = \{a_{ij}\}, i, j \in S$
Các xác suất phát ra biểu tượng	$B = \{a_{ijk}\}, i, j \in S, k \in K$
Dãy trạng thái	$X = (X_1, \ldots, X_{T+1}) \rightarrow \{1, \ldots, N\}$
Dãy output	$O = (O_1, \ldots, O_T) \quad o_t \in K$

Với một HMM cho trước, có thể dễ dàng mô phỏng hoạt động của một quá trình Markov và tạo ra một dãy output.

❖ **Các mô hình N-Gram**

N-Gram [26], [19] là một loại mô hình Markov ẩn được dùng trong các giải thuật xử lý ngôn ngữ tự nhiên sử dụng những phương pháp thống kê.

Giả sử ta cần tính xác suất xuất hiện của chuỗi T1, T2, ..., TQ

Theo công thức tính xác suất đồng thời, ta có:

$$P(T_1 T_2) = P(T_1) P\left(\frac{T_2}{T_1}\right)$$

$$P(T_1 T_2 T_3) = P(T_1 T_2) P\left(\frac{T_3}{T_1 T_2}\right)$$

Tổng quát ta có:

$$P(T_1T_2 \dots T_Q) = P(T_1)P\left(\frac{T_2}{T_1}\right)P\left(\frac{T_3}{T_1T_2}\right) \dots P\left(\frac{T_Q}{T_1T_2 \dots T_{Q-1}}\right)$$

Như vậy để tính xác suất của chuỗi $T_1T_2 \dots T_Q$, ta cần tính các xác suất điều kiện trong vế phải của công thức trên. Khi áp dụng vào bài toán xử lý ngôn ngữ tự nhiên để thống kê từ vựng, $T_1T_2 \dots T_Q$ được xem là sự xuất hiện của một từ tiếng Việt có Q tiếng. Tổ hợp sinh ra sẽ quá lớn và không thể thực hiện được nếu số tiếng tối đa trong từ là không biết trước. Tuy nhiên theo nghiên cứu của nhóm tác giả Mai Ngọc Chữ, Vũ Đức Nghiệu, Hoàng Trọng Phiến trong [8], từ tiếng Việt có tối đa 4 tiếng. Ví dụ: chủ nghĩa xã hội, đủng đa đủng đỉnh, vớ va vớ vẩn, v.v... nên hoàn toàn có thể áp dụng mô hình này.

Gọi *n* là số tiếng trong từ, công thức ước lượng xác suất dùng N-Gram như sau:

$$P\left(\frac{T_j}{T_1T_2 \dots T_{j-1}}\right) \cong \left(\frac{T_j}{T_{j-n+1}T_{j-n+2} \dots T_{j-1}}\right)$$

Có nghĩa là sự xuất hiện của tiếng thứ j chỉ phụ thuộc vào n -1 tiếng đứng trước. Thực chất N-Gram là mô hình Markov ẩn, trong đó giả sử rằng xác suất của một quan sát chỉ phụ thuộc vào n-1 quan sát trước nó.

Trong bài toán này, chúng ta sử dụng sự phân lớp của các tiếng đứng trước, sử dụng *lịch sử* để tiên đoán cho tiếng xuất hiện tiếp theo. Tuy nhiên, chúng ta không thể xem xét từng lịch sử một cách riêng biệt, vì đại đa số các câu trong ngôn ngữ tự nhiên khi xuất hiện trong văn bản không hoàn toàn giống với các câu đã xuất hiện trước đó. Thậm chí trong trường hợp phần đầu của câu đã có trong lịch sử thì phần cuối của nó vẫn hoàn toàn mới. Vì vậy không có một lịch sử nào có thể làm căn cứ tiên đoán một cách chính xác. Do đó chúng ta cần một phương pháp để thành lập các nhóm *lịch sử tương tự* làm căn cứ cho tiếng kế tiếp. Mô hình cần lập sử dụng nguyên tắc: mọi lịch sử có cùng (n-1) tiếng được xếp vào một lớp tương đương, đây

là mô hình Markov cấp (n-1) và được gọi là mô hình N-Gram. Tiếng cuối cùng trong N-Gram là tiếng cần tiên đoán sự xuất hiện của nó.

N-Gram trong bài toán này được áp dụng với n=1(Uni-Gram), n=2 (Bi-Gram) và n=3 (Tri-Gram).

❖ **Mutual information**

Mutual information hay còn gọi là thông tin tương hỗ là một ứng dụng của lý thuyết thống kê. Thông tin này dùng để đo lượng thông tin thu được về một biến ngẫu nhiên thông qua giá trị của một biến ngẫu nhiên khác. Trong xử lý ngôn ngữ tự nhiên mutual information có thể được dùng để xác định sự gắn kết giữa các âm tiết, làm cơ sở xác định các từ, các cụm từ.

Trong luận văn này, chúng tôi sử dụng mutual informati để phát hiện các từ mới (từ chưa có trong từ điển), chúng tôi sử dụng mutual information theo công thức được đưa ra bởi Zhang, J., et al in [13]:

Mutual information MI(x,y) của một Bi-Gram (x, y) được tính như sau:

$$MI(x,y) = \frac{f(x,y)}{f(x)+f(y)-f(x,y)}$$

Mutual information MI(x,y,z) của mộtTri-Gram (x,y,z) được tính như sau:

$$MI(x,y,z) = \frac{f(x,y,z)}{f(x)+f(y)+f(z)-f(x,y,z)}$$

Với:

- f(x), f(y)và f(z) là tần số xuất hiện của Uni-Gram(x), Uni-Gram(y) vàUni-Gram(z) trong ngữ liệu huấn luyện.
- f(x,y) là tần số xuất hiện củaBi-Gram(x,y) trong ngữ liệu huấn luyện.
- f(x,y,z) là tần số xuất hiện của Tri-Gram (x,y,z) trong ngữ liệu huấn luyện.

MộtN-Gramđược xác định là một từkhi mutual information >= *threshold*. Với *threshold* là giá trị ngưỡng, giá trị ngưỡng này có thể khác nhau tùy thuộc vào ngữ liệu. Trong thực nghiệm này, chúng tôi tìm được*threshold*là0.03 choBi-Gramvà0.02choTri-Gram.

CHƯƠNG 3. GIỚI THIỆU MÔ HÌNH MMSEG

3.1 Tổng quan về MMSeg

MMSeg là một hệ thống phân đoạn từ tiếng Hoa được đề xuất bởi Chih-Hao Tsai[6].Đây là hệ thống phân đoạn từ tiếng Hoa sử dụng hai dạng của thuật toán so khớp Maximum Matching kết hợp với từ điển và bốn luậtgiúp phân giải nhập nhằng. Theo kết quả tác giả đã công bố, khi thực nghiệm trên kho ngữ liệu chứa 1.013 từ, hệ thống cho kết quả rất khả quan (98.41%). Cụ thể bốn luật phân giải nhập nhằng được mô tả như sau:

- Luật số 1 : Maximum matching - so khớp tối đa:
  - Áp dụng thuật toán Maximum matching dạng đơn giản: lấy từ có chiều dài lớn nhất.
  - Áp dụng thuật toán Maximum matching dạng phức tạp: lấy từ đầu tiên từ bộ 3 từ có chiều dài lớn nhất. Nếu có nhiều hơn một bộ 3 từ có chiều dài lớn nhất thì áp dụng luật tiếp theo.
- Luật số 2 : Độ dài trung bình của từ lớn nhất:
  - Lấy từ đầu tiên của bộ 3 từ có độ dài trung bình từ lớn nhất.
  - Nếu có hơn 1 bộ 3 từ có độ dài trung bình từ lớn nhất thì áp dụng các quy tắc tiếp theo.

  Ví dụ:

  - Trường hợp 1 (TH1): _C1_C2_C3_
  - Trường hợp 2 (TH2): _C1C2C3_

  Theo luật số 2, ta lấy từ C1C2C3 ở TH2.

- Luật số 3 : độ biến đổi nhỏ nhất của chiều dài từ:
  - Chọn từ đầu tiên trong bộ 3 từ cóđộ biến đổi nhỏ nhất của chiều dài từ.
  - Nếu có nhiều hơn mộtbộ 3 từ cóđộ biến đổi nhỏ nhất của chiều dài từ thì áp dụng các quy tắc tiếp theo.

Ví dụ:

- TH1: _C1C2_C3C4_C5C6_
- TH2: _C1C2C3_C4_C5C6_

Theo luật số 3, ta lấy C1C2 trong TH1.

- Luật số 4 : tổng lớn nhất của độ tự do hình vị của các từ đơn:
  - Công thức được sử dụng để tính toán tổng độ tự do hình vị là tổng logarit tần số các từ đơn trong bộ.
  - Luật số 4 cho phép lấy từ đầu tiên của bộ có tổng logarit tần số lớn nhất.
  - Hiếm khi có trường hợp có 2 từ đơn có cùng độ tự do hình vị nên theo tác giả đến đây nhập nhằng xem như được giải quyết.

Ví dụ:

- TH1: _C1_C2_C3C4_
- TH2: _C1_C2C3_C4_

3.2 Áp dụng MMSeg vào tiếng Việt

Trong phần này chúng tôi áp dụng MMSeg trên ngôn ngữ tiếng Việt để đánh giá mức độ hiệu quả của hệ thống đối với tiếng Việt.Thực nghiệm được tiến hành trên kho ngữ liệu gồm 10.000 câu trích từ VietTreebank[22]. Ngữ liệu được sắp xếp ngẫu nhiêu sau đó chia thành 5 phần cho 5 lần thực nghiệm.

Để đánh giá, chúng tôi sử dụng các tham sốPrecision(1), Recall(2) và F-measure (3) được tính theo các công thức sau:

ức sau:

$$Precision = \frac{CorrectWords}{FoundWords} \quad (1)$$

$$Recall = \frac{CorrectWords}{StandardWords} \quad (2)$$

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision} \quad (3)$$

Với:

- StandardWords: số từ chuẩn theo ngữ liệu.
- CorrectWords: số từ phân đoạn đúng so với số từ chuẩn
- FoundWords: số từ do hệ thống tìm được.

Sau đây là kết quả thu được:

Bảng 3.1 Bảng liệt kê kết quả thực nghiệm MMSeg trên ngữ liệu tiếng Việt

Lần thực nghiệm	**Số từ chuẩn**	**Số từ tìm được**	**Số từ đúng**	**Precision(%)**	**Recall(%)**	**F-measure(%)**
		MMS	**MMS**	**MMS**	**MMS**	**MMS**
Lần 1	44,030	43,030	38,538	89.56	87.53	88.53
Lần 2	38,724	38,004	33,440	87.99	86.35	87.16
Lần 3	35,570	34,823	30,479	87.53	85.69	86.6
Lần 4	32,668	32,035	29,210	91.18	89.41	90.29
Lần 5	31,581	30,925	27,744	89.71	87.85	88.77
Trung bình	**36,515**	**35,763**	**32,282**	**89.19**	**87.37**	**88.27**

3.3 Đánh giá MMSeg trên ngôn ngữ tiếng Việt

Căn cứ vào kết quả thực nghiệm, chúng tôi nhận thấy MMSeg thu được kết quả tốt khi áp dụng trên ngôn ngữ tiếng Hoa (F-measure: 98.41%) nhưng thấp hơn nhiều khi áp dụng trên tiếng Việt (F-measure: 88.27%). Sau đây là một số nguyên nhân chúng tôi tìm được:

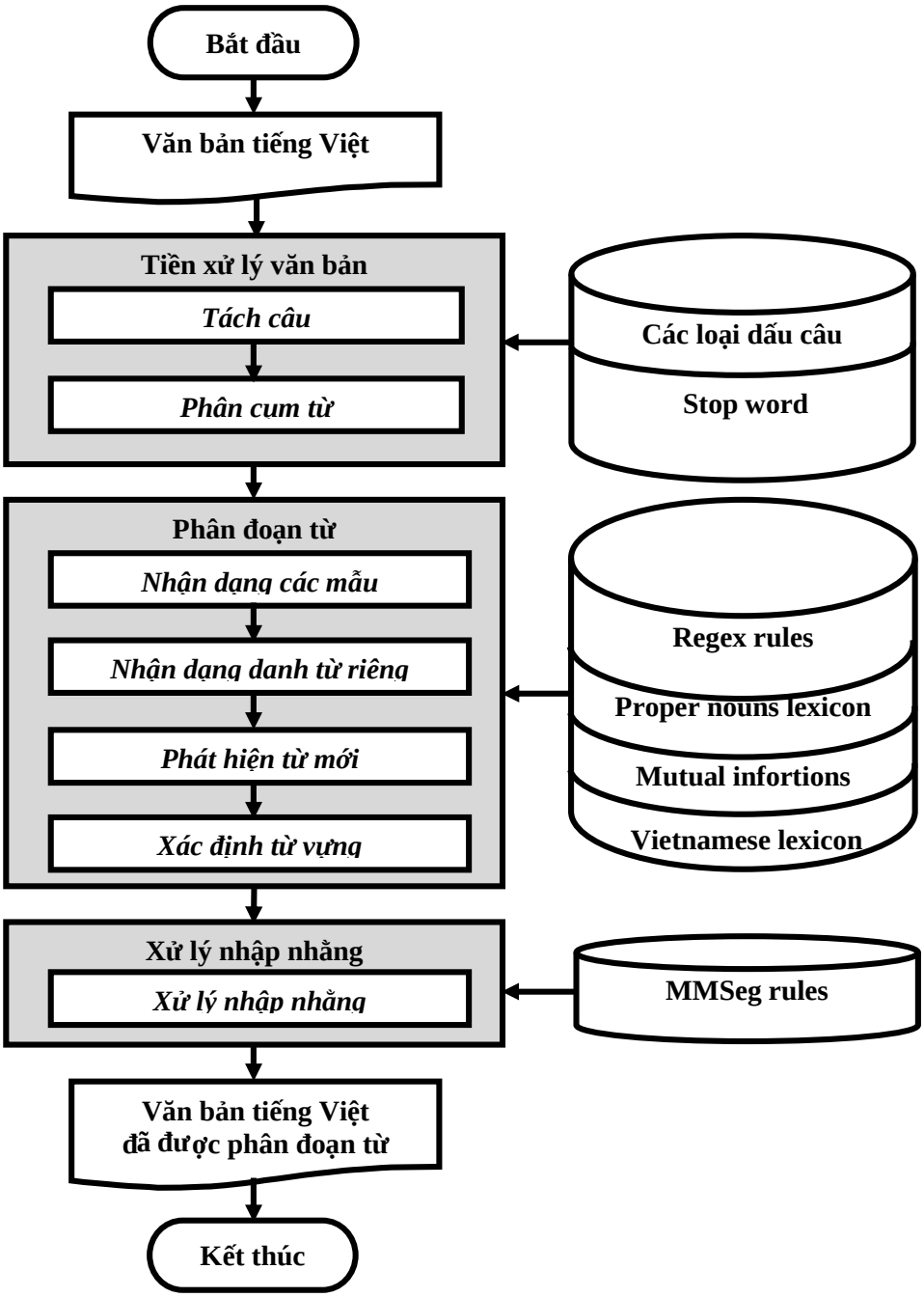
- Từ điển tiếng Việt không đầy đủ, từ mới phát sinh, từ mượn ởtiếng nước ngoài tương đối nhiều nhưng chưa được cập nhật trong từ điển.
- Các danh từ riêng như tên, địa danh, ... thường được sử dụng rất phổ biến nhưng khó để được liệt kê tất cả trong từ điển.

- Nhiều tên riêng có nguồn gốc từ danh từ (Hoa, Tùng, Lan, Ngân Hà, …), tính từ (Thương Mến, Tuấn Tú, Lung Linh, …), hoặc cụm từ (Hai Bà Trưng, Phố Hàng Bông, ...) v.v... gây ra nhiều nhập nhằng khi phân đoạn.
- Các mẫu đặc biệt như số thập phân (một phần tư, ba phần tám, ...), ngày tháng (tháng chạp, tháng giêng, ...), v.v... thường không được phân đoạn chính xác.
- Ngoài ra, việc kết hợp các tiếng để tạo nên từ trong tiếng Việt rất đa dạng cũng gây nhiều khó khăn khi phân đoạn từ.

CHƯƠNG 4. MÔ HÌNH ĐỀ XUẤT

4.1 Mô hình phân đoạn từ

Trong phần này chúng tôi đề xuất một mô hình phân đoạn từ tiếng Việt đặt tên là Vietnamese Segmention(VNS) dựa trên mô hình lai.



Hình 4.1 Hình minh họa mô hình phân đoạn từ tiếng Việt (VNS)

Mô hình được xây dựng bằng việc kết hợp nhiều phương pháp đã từng được nghiên cứu và áp dụng trước đó như: mô hình so khớp dựa trên từ điển, mô hình thống kê sử dụng N-Gram, mô hình nhận dạng từ mới sử dụng thông tin tương hỗ, mô hình nhận dạng danh từ riêng sử dụng từ điển và mô hình so trùng mẫu sử dụng các luật. Ngoài ra chúng tôi còn tăng cường thêm khả năng phân giải nhập nhằng bằng cách sử dụng giải thuật và bốn luật phân giải nhập nhằng của hệ thống MMSeg.

4.2 Thiết kế giải thuật

Đầu vào của mô hình là một văn bản tiếng Việt, đầu ra là văn bản tiếng Việt đã được phân đoạn từ. Mô hình gồm có ba bước xử lý chính. Bước 1: tiền xử lý văn bản. Bước 2: phân đoạn từ. Bước 3: xử lý nhập nhằng. Trong khuôn khổ luận văn này, chúng tôi chỉ xây dựng thuật toán cho bước 1 và bước 2. Thuật toán phân giải nhập nhằng ở bước 3 chúng tôi không xây dựng mới mà sử dụng lại thuật toán của hệ thống MMSeg trong [6].Phần tiếp theo sẽ trình bày các giải thuật.

4.2.1 Giải thuật tiền xử lý văn bản

Input:

Văn bản tiếng Việt cần phân đoạn từ, danh sách dấu câu, danh sách stop word.

Output:

Danh sách các cụm từ.

Giải thuật:

```
1.Open file VanBan.txtas Text for"Read"
2.Open fileDSDauCau.txtas SignList for"Read"
3.Open fileDSStopWord.txtas StopWordList for"Read"
4.SentenceList  =[]// Chứa danh sách các câu
5.ChunkingList =[]// Chứa danh sách các cụm từ
6.foreach(sign in SignList)do
7.Sentence= DoSentenceChunking(Text, sign)
8.addSentencetoSentenceList
9.end
10.for each (sentence  in SentenceList )do
11.for each (stopWord in StopWordList)do
12.Chunking= DoPhraseChunking(sentence, sign)
13.add Chunking to ChunkingList
14.end
15.end
16.return ChunkingList
```

Trong giải thuật này chúng tôi tiến hành phân đoạn câu và phân đoạn cụm từ dựa trên các dấu câu và cácstop word. Mục đích nhằm rút ngắn thời gian xử lý cho các bước tiếp theo.

Dấu câu là các loại dấu câu thường dùng trong tiếng Việt như: dấu chấm, dấu phảy, dấu chấm phảy, dấu chấm than, v.v...

Stop word là những từ thường xuất hiện trong hầu hết các tài liệu, thường dùng để liên kết các mệnh đề, các từ và cụm từ với nhau. Luận văn sử dụng danh sách stop word được thống kê bởi tác giả Nguyễn Đức Dân trong [9] để phân cụm từ. Sau đây là danh sách một số stop word luận văn sử dụng:

Bảng 4.1 Danh sách một số stop word trong tiếng Việt

và	còn	hay	hoặc	không
không những	không chỉ	mà	còn	nếu
thì	nên	hễ	tuy	nhưng
vả lại	giá	vì	bởi	tại
do	song	dầu	mặc dầu	dù
dẫu	dẫu cho	chẳng lẽ	làm như	thế mà
vậy mà	có điều	hơn nữa	huống hồ	huống gì
huống nữa	ngay	cũng	chính	cả

4.2.2 Giải thuật phân đoạn từ

Input:danh sách các cụm từ (*ChunkingList*), danh sách biểu thức chính quy, từ điển danh từ riêng, thông tin tương hỗ, từ điển tiếng Việt

Output:danh sách các từ

Giải thuật:

```
1.Open file BieuThucChinhQuy.txt as RegexRuleList for"Read"
2.Open file TuDienDanhTu.txt as ProperNounDic for"Read"
3.Open file ThongTinTuongHo.txt as MutualInforDic for"Read"
4.Open file TuDienTiengViet.txt asVNLexicon for"Read"
5. WordList  =[]// Chứa danh sách các từ
6.for each (chunk in ChunkingList)do
7.wordsMatch1= DoMatchingRegex(chunk, RegexRuleList, ref chunk2)
8.wordsMatch2= DoMatchingProperNoun(chunk2, ProperNounDic, ref chunk3)
9.wordsMatch3= DoDetectNewWord(chunk3, MutualInforDic, ref chunk4)
10.wordsMatch4= DoDetectWord(chunk4, VNLexicon)
11.addwordsMatch1,wordsMatch2,wordsMatch3,wordsMatch4 toWordList
10.end
11.return WordList;
```

Mô tả giải thuật

Đầu vào của giải thuật là danh sách các cụm từ và các thông tin tham khảo, tiến hành tạo một danh sách các từ gợi ý theo các bước:

- Bước 1: tạo một danh sách các từ gợi ý sử dụng phương pháp so trùng mẫu, với các luật được định nghĩa bằng các biểu thức chính quy.
- Bước 2: tạo một danh sách các từ gợi ý sử dụng từ điển danh từ riêng.
- Bước 3: tạo một danh sách các từ gợi ý sử dụng từ điển từ vựng.
- Bước 4: tạo một danh sách các từ gợi ý sử dụng thông tin tương hỗ dựa trên ngữ liệu huấn luyện N-Gram, giúp xác định các từ mới.

Đầu ra của giải thuật là một danh sách các từ. Danh sách này sẽ được dùng như từ điển từ vựng phục vụ cho giải thuật phân giải nhập nhằng.

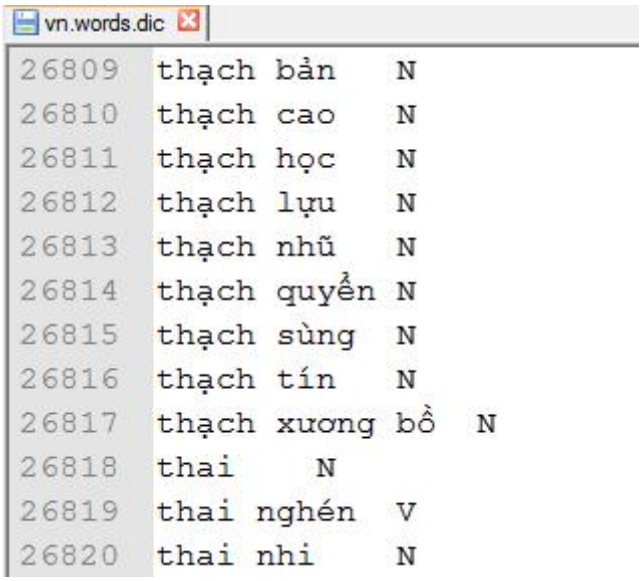
Phần tiếp theo chúng tôi trình bày về quá trình thu thập và xây dựng các từ điển phục vụ cho thực nghiệm.

4.2 Từ điển và kho ngữ liệu

Để thực nghiệm, chúng tôi tiến hành thu thập và xây dựng từ điển từ vựng, từ điển danh từ riêng, kho ngữ liệu để huấn luyện N-Gram và tập luật nhận dạng các mẫu. Sau đây là kết quả thu được:

- **Từ điển từ vựng**

Từ điển từ vựng luận văn này sử dụng là từ điển tiếng Việt - Vietnamese Lexicon [16], một sản phẩm của dự án KC01.01/06-10, đã tùy biến lại cấu trúc cho phù hợp. Phiên bản luận văn sử dụng có 3,1243 từ.



Hình 4.2 Hình minh họa cấu trúc từ điển tiếng Việt

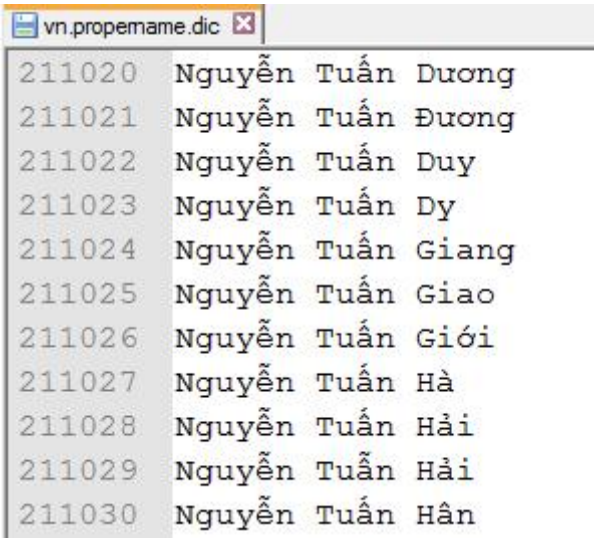
❖ **Từ điển danh từ riêng**

Theo khảo sát của chúng tôi, trong khuôn khổ đề tài cấp nhà nước KC.01.21 – "Nghiên cứu các kỹ thuật xây dựng và khai thác thông tin Web có ngữ nghĩa", hiện nhómtác giả Cao Hoàng Trụ đã và đang xây dựng một cơ sở tri thức rất lớn về các thực thể có tên tiếng Việt. Dự án này có tên là VN-KIM [25]. VN-KIM Ontology bao gồm 347 lớp thực thể và 114 quan hệ và thuộc tính. Với khoảng 85,767 tên đối tượng.

VN-Kim thực sự rất phù hợp làm từ điển danh từ riêng trong mô hình đề xuất. Tuy nhiên, vì chưa thể liên hệ được với tác giả để xin phép sử dụng dữ liệu nên chúng tôi tự xây dựng từ điển danh từ riêng phục vụ cho việc thực nghiệm. Từ điển có hai loại: tên người và tên địa danh. Dữ liệu được thu thậptự động từ danh bạ điện thoại trực tuyến *nhungtrangvang.com.vn* và bách khoa toàn thư*vi.wikipedia.org*. Sau khi loại bỏ các tên trùng nhau, các tên không hợp lệ chúng tôi thu được từ điển gồm 1,068,435 tên. Cụ thể:

Bảng4.2 Bảng liệt kê số lượng têncủa từ điển danh từ riêng

Loại tên	Số lượng
Tên người	1,065,613
Tên địa danh	2,822
Tổng cộng	1,068,435



Hình 4.3 Hình minh họa từ điển danh từ riêng

❖ **Ngữ liệu để huấn luyện N-Gram**

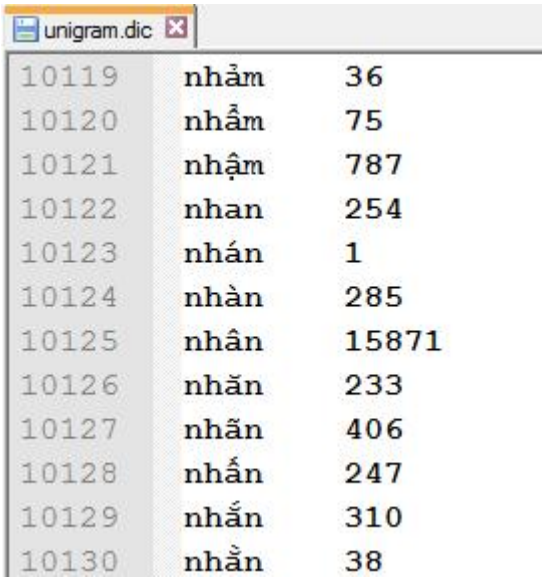
Ngữ liệu để huấn luyện cho mô hình N-Gram được chúng tôi thu thập tự động từ các báo điện tử trực tuyến như: vietnamnet.vn, vnexpress.net, dantri.com.vn, tuoitre.com.vn, vnthuquan.net, v.v... Trước khi huấn luyện chúng tôi làm sạch văn bản bằng cách: bóc tách toàn bộ html, giữ lại phần nội dung chính, xóa các nội dung mang tính lặp lại như: menu, logo, quảng cáo, các liên kết liên quan, ... Tổng dung lượng thu được là 26MB với 3356 bài báo. Cụ thể:

Bảng4.3 Bảng liệt kê số lượng bài báo phục vụ cho việc huấn luyện dữ liệu

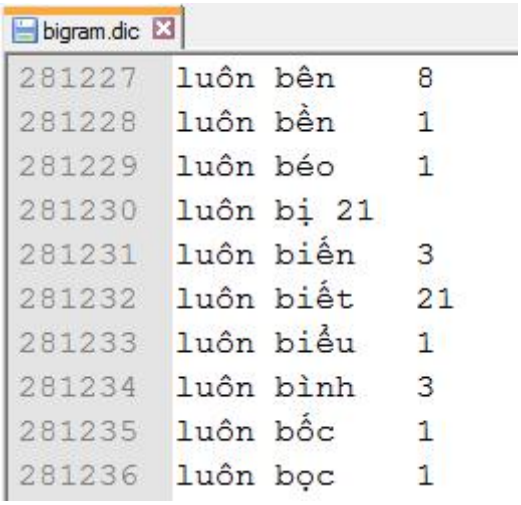
Loại bài báo	Số lượng
Đời sống	641
Giải trí	367
Kinh tế	340
Văn hóa	352
Xã hội	525
Công nghệ	231
Nghệ thuật	450
Thể thao	316
Sự kiện	134
Tổng cộng	3356

Sau khi huấn luyện chúng tôi thu được tậpngữ liệu huấn luyện gồm có:

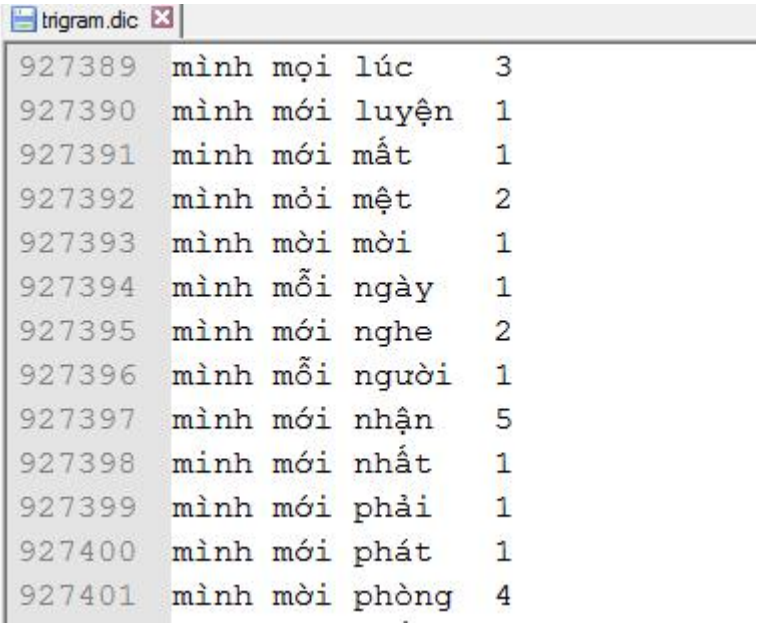
- Uni-Gram: 16.340 mục
- Bi-Gram: 571.821 mục
- Tri-Gram: 1.769.290 mục



Hình 4.4 Hình minh họa kết quả huấn luyện Uni-Gram



Hình 4.5 Hình minh họa kết quả huấn luyện Bi-Gram



Hình 4.6 Hình minh họa kết quả huấn luyện Tri-Gram

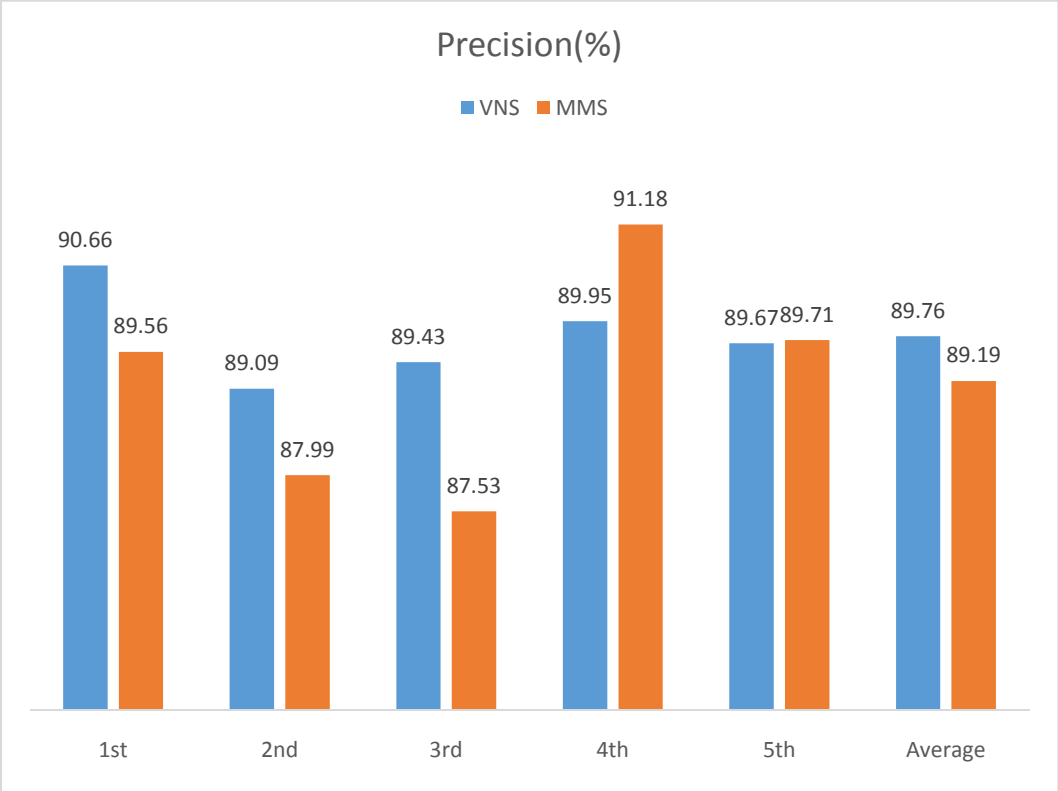
4.3 Thực nghiệm

Thực nghiệm được tiến hành trên các mẫu ngữ liệu đã dùng trong mục 5.2 với cùng một phương pháp đánh giá nhằm so sánh hai mô hình.

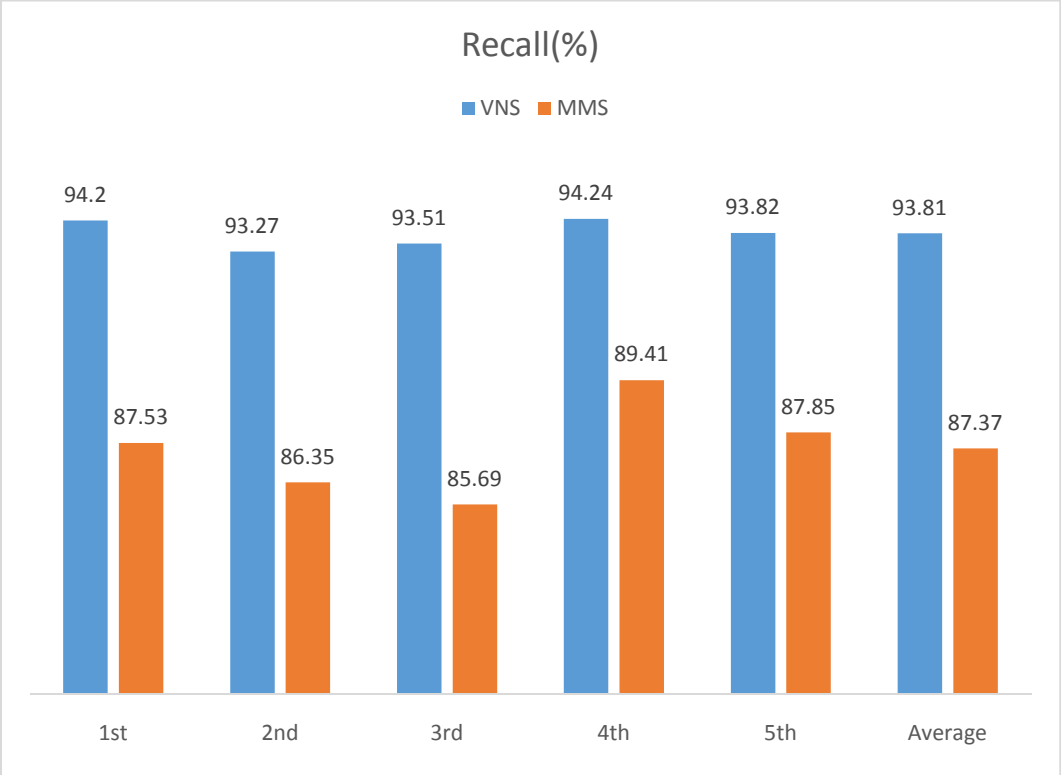
Bảng 4.4 trình bày kết quả của mô hình do chúng tôi đề xuất(VNS) so với mô hình củaMMSeg(MMS)sau quá trình chạy thực nghiệm:

Bảng 4.4 Bảng liệt kê kết quả thực nghiệm của VNS so với MMS

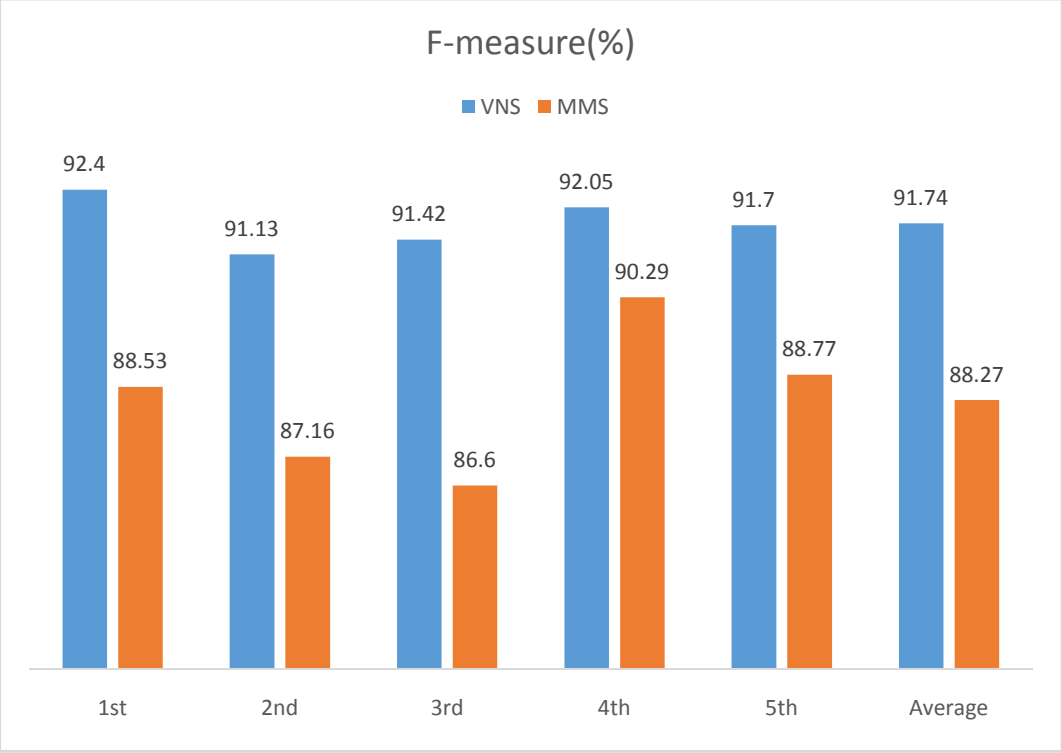
Lần thực nghiệm	Số từ chuẩn	Số từ tìm được		Số từ đúng		Precision(%)		Recall(%)		F-measure(%)	
		VNS	MMS	VNS	MMS	VNS	MMS	VNS	MMS	VNS	MMS
1	44,030	45,751	43,030	41,477	38,538	90.66	89.56	94.2	87.53	92.4	88.53
2	38,724	40,541	38,004	36,117	33,440	89.09	87.99	93.27	86.35	91.13	87.16
3	35,570	37,195	34,823	33,263	30,479	89.43	87.53	93.51	85.69	91.42	86.6
4	32,668	34,225	32,035	30,787	29,210	89.95	91.18	94.24	89.41	92.05	90.29
5	31,581	33,041	30,925	29,629	27,744	89.67	89.71	93.82	87.85	91.7	88.77
Trung bình	**36,515**	**38,151**	**35,763**	**34,255**	**32,282**	**89.76**	**89.19**	**93.81**	**87.37**	**91.74**	**88.27**



Hình4.7So sánhtham số Precision của mô hình VNS và MMS



Hình 4.8 So sánh tham số Recall của mô hình VNS và MMS



Hình 4.9 So sánh tham số F-Measure của mô hình VNS và MMS

Kết quả sau năm lần thực nghiệm, về tổng quancác tham số Precision, Recall và F-measure trung bình của mô hình đề xuất VNS đều cao hơn so với mô hình củaMMS. Tuy nhiên, nếu xét riêng tham số Precision, cũng có những trường hợp độ chính xác Precision của VNS thấp hơn MMS. Cụ thể là ở lần thực nghiệm thứ 4: VNS: 89,95 – MMS: 91,18; ở lần thực nghiệm thứ 5: VNS: 89,67 – MMS: 89,71.

Qua khảo sát dữ liệu, chúng tôi nhận thấyVNS tỏ ra hiệu quả hơn khi xử lý những văn bản có chứa nhiều danh từ riêng, văn bản có chứa những từ chưa có trong từ điển, nhưng có xuất hiện thường xuyên trong tập huấn luyện, văn bản có chứa các mẫu đặc biệt. Từ đó phần nào thấy được những cải tiến trong mô hình của VNS so với mô hình của MMS là có hiệu quả.

CHƯƠNG 5. KẾT LUẬN

5.1 Nhận xét chung

Trong xử lý ngôn ngữ tự nhiên tiếng Việt, phân đoạn từ là một trong những công việc hết sức quan trọng. Do tính chất đặc biệt của ngôn ngữ tiếng Việt, công việc này có tính khó cao và độ phức tạp lớn. Giải quyết tốt bài toán này sẽ làm tiền đề vững chắc cho các bài toán lớn hơn như: dịch máy tự động, tóm lược văn bản, xử lý ngữ nghĩa, v.v...

Đã có nhiều mô hình phân đoạn từ được đưa ra bởi nhiều nhà nghiên cứu trước đó. Mỗi phương pháp đều có ưu và nhược điểm riêng.Nhìn chung việc phân đoạn từđều chưa thể đạt mức tyệt đối chính xác, ngay cả khi con người trực tiếp thực hiện [21]. Nguyên nhân một phần là do khả năng xử lý ngôn ngữ trên máy tính của con người còn hạn chế, một phần khácla do bản thân ngôn ngữ luôn vận động và phát triển không ngừng theo thời gian, đòi hỏi người làm xử lý ngôn ngữ phải luôn nghiên cứu và cải tiến phương pháp.

Trong các mô hình phân đoạn từ mà luận văn tham khảo được, mô hình sử dụng từ điển kết hợp với các luật để phân giải nhập nhằng của hệ thống MMSeg là khá phù hợp với ngôn ngữ tiếng Việt nên chúng tôi chọn làm hướng tham khảo chính. Tuy nhiên, khi áp dụng trên ngôn ngữ tiếng Việt, với những đặc thù riêng, MMSeg đã không thực sự hiệu quả.

Qua quá trình nghiên cứu và thực nghiệm, luận văn đã xây dựng được một mô hình phân đoạn từ mới, dựa trên mô hình lai với cơ sở là bốn luật phân giải nhập nhằng của hệ thống MMSeg kết hợp với các phương pháp khác như: tự động phát hiện từ mới, nhận dạng danh từ riêng, so trùng mẫu, phương pháp thống kê.

Kết quả thực nghiệm đã chứng tỏ mô hình mới có đem lại một số kết quả khả quan, là động lực cho những nghiên cứu tiếp theo.

5.2 Kết quả đạt được

Trong khuôn khổ luận văn cao học ngành Công nghệ thông tin, luận văn đã nghiên cứu và đưa ra một mô hình phân đoạn từ tiếng Việt. Bước đầu luận văn đã thu được một số kết quả sau:

- Xây dựng được một từ điển danh từ riêng với 1.068.435 tên. Trong đó có 1,065,613 tên người và 2,822 tên địa danh ở Việt Nam.
- Xây dựng được một kho ngữ liệu gồm có 3.356 bài báo chứa trong 26Mb dữ liệu.
- Thực nghiệm và đánh giá hệ thống MMSeg trên ngôn ngữ tiếng Việt với ngữ liệu gồm 10,000 câu trích từ VietTreebank.
- Xây dựng một mô hình phân đoạn từ cho tiếng Việtdựa trên mô hình lai sử dụng nhiều phương pháp tích hợp: phương pháp dựa trên từ điển, phương pháp dựa trên thống kê, phương pháp so trùng mẫu và phương pháp phân giải nhập nhằng dựa trên các luật.

5.3 Hạn chế của đề tài

Vì thời gian có hạn, nên đề tài còn hạn chế trong một số vấn đề sau:

- Ngữ liệu huấn luyện cho mô hình N-Gram còn nhỏ, chỉ giới hạn ở một số lĩnh vực nhất định, không mang tính phổ quát, phần nào ảnh hưởng đến độ chính xác của mô hình.
- Từ điển danh từ riêng xây dựng được còn phải bổ sung thêm nhiều.
- Số lần thực nghiệm và dữ liệu thực nghiệm còn ít nên phần đánh giá còn mang tính chủ quan.

Tuy đã nỗ lực thực hiện, nhưng chắc chắn đề tài vẫn còn nhiều thiếu sót, em rất mong tiếp tục nhận được sự chỉ dẫn của quý thầy cô.

5.4 Hướng phát triển của đề tài

Qua quá trình khảo sát và thực nghiệm, chúng tôi nhận thấy bốn luật phân giải nhập nhằng của hệ thống MMSeg mà mô hình đang áp dụng là chưa đủ cho ngôn ngữ tiếng Việt. Các phân đoạn sai do nhập nhằng còn rất nhiều và phức tạp. Các luật hiện tại chủ yếu chỉ xử lý nhập nhằng dựa trên xác xuất của từ và cụm từ.

Trong tương lai, chúng tôi muốn tăng cường thêm các luật phân giải nhập nhằng mới có xét đến khía cạnh ngữ nghĩa và ngữ pháp. Cụ thể, trong những nỗ lực tiếp theo, chúng tôi muốn áp dụng thêm cây phân tích cú pháp VietTreebank vào quá trình phân giải nhập nhằng, nhằm nâng cao hơn độ chính xác khi phân đoạn từ.

TÀI LIỆU THAM KHẢO

[1] Trần Ngọc Anh, Nguyễn Nhật An. (2011). Lựa chọn tập gán nhãn ranh giới từ cho mô hình Markov ẩn trong bài toán tách từ tiếng Việt.

[2] Luu Tuan Anh, Yamamoto Kazuhide. (2012). A pointwise approach for Vietnamese Diacritics Restoration. 2012 International Conference on Asian Language Processing, pp.189 – 192.

[3] Ngoc Anh Tran, Thanh Tinh Dao, Phuong Thai Nguyen. (2012). An effective context-based method forVietnamese-word segmentation. IEEE 9th, pp.34-40.

[4] Dương Hữu Biên. (2010). Giáo trình cơ sở ngôn ngữ học, ĐH Đà Lạt.

[5] Nguyễn Tài Cẩn. (1975). Ngữ pháp tiếng Việt, Tiếng - Từ ghép - Đoản ngữ, Nxb Khoa học xã hội, Hà Nội.

[6] Chih-Hao Tsai. (1996). MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm.www.casper.beckman.uiuc.edu/~ctsai4/chinese/wordseg/mmseg.html.

[7] Mai Ngọc Chừ. (1997). Cơ sở ngôn ngữ học và tiếng Việt, Nxb Giáo dục, trang 91–105.

[8] Mai Ngọc Chừ, Vũ Đức Nghiệu và Hoàng Trọng Phiến. (1997). Cơ sở ngôn ngữ học và tiếng Việt. Nxb Giáo dục, trang 142–152.

[9] Nguyễn Đức Dân. (1987). Lôgic ngữ nghĩa cú pháp. NXB ĐH&TH chuyên nghiệp, Hà Nội.

[10] Đinh Điền, Hồ Bảo Quốc. (2008). Vấn đề về ranh giới từ trong ngữ liệu song ngữ Anh-Việt

[11] Nguyễn Thiện Giáp. (1998). Dẫn luận Ngôn ngữ học, Nxb Giáo dục, trang 298–305.

[12] Cao Xuân Hạo. (2003). Tiếng Việt - Mấy vấn đề Ngữ âm, Ngữ pháp, Ngữ nghĩa. Nxb Khoa học xã hội.

[13] Lê Trung Hiếu, Lê Anh Vũ, Lê Trung Kiên. (2013). Áp dụng xác suất thống kê và quá trình máy tự học cho bài toán phân tách từ văn bản tiếng Việt. Tạp chí Khoa học & Công nghệ Đại học Duy Tân số 6, trang 32-38.

[14] Hla Hla Htay, Kavi Narayana Murthy. (2008). Myanmar Word Segmentation using Syllable level Longest Matching. Proceedings of the 6th Workshop on Asian Language Resources, pp.41-48.

[15] H. P. Lê, T. M. H. Nguyen, A. Roussanaly and T. V. Ho. (2008). A hybrid approach to word segmentation of Vietnamese texts. In 2nd International Conference on Language and Automata Theory and Applications, Tarragona, Spain, pp.240-249.

[16] Thi Minh Huyen Nguyen, Laurent Romary, Mathias Rossignol, Xuan Luong Vu. (2006). A lexicon for Vietnamese language processing. Language Reseourse Evaluation - Volume 40, pp.291-309.

[17] Nguyễn Thị Minh Huyền, Hoàng Thị Tuyền Linh, Vũ Xuân Lương. (2009). Hướng dẫn nhận diện đơn vị từ trong văn bản tiếng Việt.

[18] Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo. (2005). A Maximum Entropy Approach to Chinese Word Segmentation. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pp.161-164.

[19] Jurafsky and Martin. (2009). Speech and Language Processing: An Introduction to Speech Recognition. Computational Linguistics and Natural Language Processing, SE, Prentice Hall, pp.934.

[20] Masaaki Nagata. (1997). A self-organizing Japanese word segmenter using heuristic word identication and re-estimation. In Joe Zhou and Kenneth Church, editors, Proceedings of the Fifth Workshop on Very Large Corpora, pp.203-215.

[21] Richard Sproat, Chilin Shih, William Gale, Nancy Chang. (1994). A stochastic finite-state word-segmentation algorithm for Chinese. ACL '94 Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp.66-73.

[22] Phuong-Thai Nguyen, Xuan-Luong Vu, Thi-Minh-Huyen Nguyen, Van-Hiep Nguyen, Hong-Phuong Le. (2009). Building a large syntactically-annotated corpus of Vietnamese. Proceedings of the Third Linguistic Annotation Workshop, Suntec, Singapore, pp.182-185.

[23] Nguyễn Kim Thản. (1997). Nghiên cứu ngữ pháp tiếng Việt. NXB GD, trang 28.

[24] Theeramunkong, T., Usanavasin, S. (2001). Non-dictionary-based Thai word segmentation using decision trees. The first international conference on Human language technology research. New Jersey, USA (2001), pp.1-5.

[25] Truc-Vien T. Nguyen., Tru H. Cao. (2007). VN-KIM IE: Automatic extraction of vietnamese named-entities on the web. New Generation Computing May 2007, Volume 25, Issue 3, pp 277-292.

[26] Trần Ngọc Tuấn. (2002). Phân đoạn từ tiếng Việt dùng Corpus và các mô hình thống kê, luận văn thạc sỹ, Đại học Bách Khoa TP.Hồ Chí Minh.