Hệ thống tổng hợp tiếng nói tiếng Việt

Nguyễn Trọng Hiếu, Lê Quang Thắng, Lê Anh Tú, Đỗ Văn Thảo, Nguyễn Hữu Thuận.

***Tóm tắt* – Tổng hợp tiếng nói là một lĩnh vực có ứng dụng rộng rãi và được rất nhiều quan tâm nghiên cứu trên thế giới cũng như ở Việt Nam. Hiện nay tại Việt Nam đã phát triển nhiều bộ tổng hợp tiếng nói dành riêng cho tiếng Việt. Tuy nhiên, chất lượng tiếng nói tổng hợp sao cho dễ nghe và tự nhiên vẫn là điều mà các nhà nghiên cứu đang hướng tới. Nghiên cứu này tập trung vào toàn bộ thành phần của một bộ tổng hợp tiếng nói, đề xuất và cài đặt các thuật toán và mô hình cho việc cải thiện chất lượng tiếng nói tổng hợp thông qua cải tiến từng thành phần trong hệ thống tổng hợp tiếng nói. Mục tiêu cuối cùng là có thể xây dựng một hệ thống tổng hợp tiếng nói hoàn chỉnh các thành phần với chất lượng tốt.**

***Từ khóa* – cao độ, chuẩn hóa, lựa chọn đơn vị, phân tích cú pháp, tổng hợp tiếng nói, trường độ.**

1. GIỚI THIỆU

Tổng hợp tiếng nói là quá trình tạo ra tiếng nói nhân tạo của người trên máy tính từ văn bản. Đây là một đề tài có tính ứng dụng thực tiễn cao nên được nghiên cứu nhiều trên thế giới và Việt Nam từ rất sớm. Ứng dụng của tổng hợp tiếng nói có thể dễ dàng thấy trong nhiều hệ thống, như hệ thống hỗ trợ đọc văn bản cho người khuyết tật, hệ thống trả lời tự động tại các tổng đài hay robot, hệ thống chỉ đường trong các phương tiện vận tải...

Hình 1 Hệ thống tổng hợp tiếng nói

Bộ tổng hợp tiếng nói được chia làm hai phần chính: tổng hợp mức cao và tổng hợp mức thấp. Nhiệm vụ phần tổng hợp mức cao là chuẩn hóa văn bản, phát sinh thông tin về ngữ âm, ngữ điệu.

Phần tổng hợp mức thấp dựa vào các thông tin phía trên sẽ tiến hành tìm kiếm và lựa chọn đơn vị âm, thực hiện ghép nối và làm trơn tín hiệu, cho ra tiếng nói cần tổng hợp.

2. CHUẨN HÓA VĂN BẢN

Trong hệ thống tổng hợp tiếng nói, việc chuẩn hóa văn bản là công đoạn đầu tiên có ảnh hưởng quan trọng trong việc đảm bảo văn bản được đọc một cách đúng đắn.

Nhóm sinh viên: Nguyễn Trọng Hiếu, Lê Quang Thắng, Lê Anh Tú, Đỗ Văn Thảo, Nguyễn Hữu Thuận, lớp Công nghệ phần mềm, , khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 01677225100, e-mail: tronghieubk@gmail.com).

Giáo viên hướng dẫn:

TS. Trần Đỗ Đạt, Trung tâm nghiên cứu Mica.

ThS. Nguyễn Thị Thu Trang, Bộ môn Công nghệ phần mềm, Viện CNTT-TT.

Hiện tại đã có một số nghiên cứu về chuẩn hóa văn bản trong tiếng Việt, nhưng kết quả chủ yếu mới chỉ dừng lại ở những tập luật cơ bản áp dụng cho những trường hợp đặc biệt, chưa giải quyết được bài toán một cách hệ thống.

Trong nghiên cứu này, chúng tôi xem xét bài toán một cách tổng quát để đưa ra giải pháp tổng thể cho việc chuẩn hóa văn bản tiếng Việt. Các vấn đề về các dạng chưa chuẩn và các tình huống nhập nhằng được giải quyết.

2.1 Non standard words

Các trường hợp cần phải chuẩn hóa được quan sát và phân loại vào các dạng khác nhau gọi là "các loại từ chưa chuẩn hóa" hay Non-standard Word (NSW). Mỗi loại từ chưa chuẩn hóa có cách xử lý riêng. Việc phân loại các từ chưa chuẩn hóa được thể hiện trong bảng sau:

Nhóm	*Loại*	*Mô tả*	*Ví dụ*
Số	NTIM	Thời gian/giờ	1:30
	NDAT	Ngày/tháng/năm	17/3/87
	NDAY	Ngày và tháng	17/3, 03-05/3
	NMON	Tháng và năm	3/88, 5/2011
	NNUM	Số/số học	2009, 70.000
	NTEL	Số điện thoại	0915.334.577
	NDIG	Số hiệu, mã số	VN534
	NSCR	Tỉ số	Tỉ số là 3-5
	NRNG	Miền giá trị	Từ 3-5 ngày
	NPER	Số phần trăm	93%, 30-40%,
	NFRC	Phân số	34/6, 6/145
	NCOM	Hỗn hợp	2x2x3, 18+, 2*3
	NADD	Địa chỉ	Ngách 128/27/2A
	NSIG	Kí hiệu	m2, m3
Chữ	LWRD	Từ mượn	London, NATO
	LSEQ	Dãy các ký tự	ODA, GDP
	GREK	Số Hi Lạp	I, II
	LABB	Viết tắt	TS (tiến sĩ)
Khác	PUNC	Dấu câu đọc được	… ( ) [ ] ‘ ’ “ ” - /
	SENT	Dấu phân tách câu	. ? ! …
	URLE	Địa chỉ url, email	http://soict.hut.vn
	MONY	Tiền tệ	2$, $2, 100 ¥,
	DURA	Trường độ (nghỉ)	“-” in scores (2-3)
	NONE	Bỏ qua	asscii art…

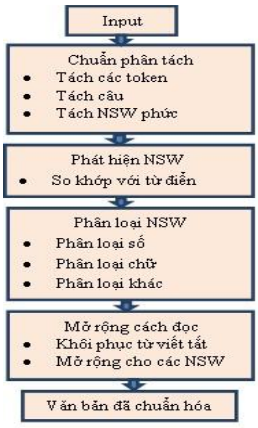
Bảng 1 Bảng phân loại NSW

Văn bản đầu vào là văn bản lấy trong thực tế, ban đầu rất hỗn độn vì nó chứa nhiều dạng từ chưa chuẩn hóa khác nhau. Vấn đề là nhận ra và phân loại đúng những từ này. Vì mỗi loại có cách đọc khác nhau nên khi phân loại sai có thể sẽ khiến cách đọc sai và người nghe hiểu sai nội dung văn bản. Ví dụ "phần XI" đọc lên là "phần mười một", nếu không nhận đúng số la mã "XI" sẽ đọc là "phần xi"!

Để đảm bảo tính nguyên vẹn của văn bản đầu vào mà vẫn bổ xung được thông tin sau quá trình chuẩn hóa, chúng tôi tổ chức lại văn bản và thông tin bổ xung dưới dạng cấu trúc XML. Các thông tin là thuộc tính của các thẻ XML, khi bỏ qua các thẻ này ta nhận đượcvăn bản gốc.

2.2 Các bước của quy trình chuẩn hóa.

Chuẩn phân tách: Văn bản đầu vào trước hết được xử lý bằng regex để nhận ra và đánh dấu các tổ hợp thuộc nhóm số, URLE bởi các nhóm này sẽ được xử lý riêng. Tiếp đó các dấu trắng thừa trong văn bản được loại bỏ, thêm dấu trắng vào trước và sau các dấu câu, các khoảng trắng trong một tổ hợp số được thay bởi dấu chấm "." để tiện cho việc xử lý về sau. Các câu trong văn bản được phân tách và đánh dấu, phục vụ cho việc khai thác ngữ cảnh và đưa ra nhịp điệu đọc phù hợp cho tiếng nói tổng hợp. Cuối cùng các NSW phức được tách ra chuẩn bị cho bước sau.



Hình 1 - Quy trình chuẩn hóa văn bản

Phát hiện NSW: Các token được tìm kiếm trong một từ điển các âm tiết tiếng Việt có thể đọc được, nếu không thấy thì đánh dấu là NSW.

Phân loại NSW: Các NSW nhóm số, nhóm chữ, nhóm khác được phân nhóm đơn giản bởi các luật nhận ra định dạng của chúng trong regex. Việc phân loại các NSW nhóm số, nhóm chữ phức tạp hơn vì tiềm ẩn nhiều trường hợp nhập nhằng nên được giải quyết bằng các kỹ thuật data mining. Các dạng trong nhóm số được phân loại bằng cây quyết định, với các thuộc tính định dạng và ngữ cảnh. Định dạng gồm số các chữ, số các số, miền giá trị, ngữ cảnh gồm 2 token liền trước và liền sau của NSW. Các dạng trong nhóm chữ không có định dạng đặc trưng để nhận ra, ngữ cảnh cũng kém rõ ràng hơn dạng số. Vì thế dạng chữ được phân loại bằng việc tính xác suất của một NSW thuộc về LWRD, LSEQ hay LABB dựa trên một tập huấn luyện. NSW được phân loại theo xác suất lớn nhất. Các trường hợp thuộc nhóm khác được phân loại dựa vào định dạng, SENT được phân loại bởi việc phân tách câu bước chuẩn phân tách.

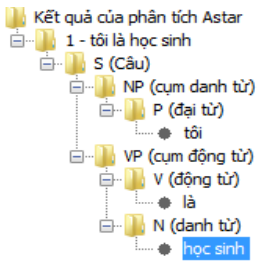
Mở rộng cách đọc: Vấn đề khôi phục từ viết tắt cũng tiềm ẩn rất nhiều nhập nhằng. Cùng một NSW có thể tìm được nhiều từ đầy đủ thỏa mãn nó. Một danh sách các từ viết tắt và các từ đầy đủ được sử dụng. Với mỗi LABB sẽ duyệt tìm các từ đầy đủ có thể của nó, sau đó dùng mô hình ngôn ngữ để khai thác ngữ cảnh và tính xác suất từ đầy đủ và entropy cho LABB đó. Tổng hợp lại ta sẽ lựa chọn lấy từ đầy đủ là trường hợp có tích xác suất và entropy lớn nhất.

Cách đọc các nhóm số đưa ra bởi các luật khá đơn giản. Từ mượn được đọc dựa vào từ điển từ mượn, từ viết tắt được đọc theo từ đầy đủ của nó, dãy chữ và dãy số được đọc từng kí tự, các dấu đọc theo cách phát âm thông thường của nó, dấu phân tách câu không được đọc.

3. PHÂN TÍCH CÚ PHÁP.

Trong tổng hợp tiếng nói, phân tích cú pháp đóng một vai trò rất quan trọng trong công đoạn xử lí văn bản của hệ thống. Phân tích cú pháp chuẩn xác sẽ đưa ra cho hệ thống một cái nhìn toàn cảnh về cấu trúc của văn bản, các cụm từ trong văn bản từ phức tạp cho đến đơn giản nhất, đồng thời các vị trí âm tiết trong cụm từ cũng được đưa ra luôn.

Mục đích của bộ phân tích cú pháp là đưa ra được cây phân tích cú pháp của văn bản đầu vào. Dưới đây là một ví dụ về cây phân tích cú pháp của một câu:



Trong phần này, chúng tôi đã nghiên cứu cách thức để có thể cải tiến đầu ra cho bộ phân tích cú pháp cả về mặt tốc độ cũng như chất lượng.

3.1 Mô hình xác suất PCFG.

Mô hình PCFG là một mô hình văn phạm phi ngữ cảnh dùng để biểu diễn và quản lí tập luật cú pháp tiếng Việt. Mô hình PCFG là một tập bao gồm 5 tham số G=(N,∑,P,S,D) [5], trong đó :

- N : tập các nhãn từ loại, $\{N^i\}$, i=1,... n
- ∑ : tập các từ được tách từ văn bản, $\{W^k\}$, k=1,..., V
- P : tập các luật có dạng $\{N^i \rightarrow \zeta^j\}$, $\zeta^j \in (\Sigma \cup N)^*$
- S : ký hiệu khởi đầu, tượng trưng cho một câu.
- D : tính xác suất cho mỗi luật tương ứng trong P .

Với PCFG, ta có xác suất của một cây phân tích cú pháp

$P(T) = \sum_{m}^{i=1} lg\ (r_i)$, [6]

trong đó r_i là các luật sử dụng trong cây

Với một câu đầu vào, sẽ có nhiều cây phân tích cú pháp đầu ra, cây nào có xác suất cao nhất sẽ là cây đầu ra thích hợp nhất. Vậy vấn đề chúng ta đặt ra ở đây là phải tìm được giải thuật giúp được cây phân tích cú pháp có xác suất lớn nhất với thời gian ngắn nhất.

3.1.1 Xác suất inside.

Nếu có một nút N^j được tạo ra bởi luật $N^j \rightarrow N^r\ N^s$ thì inside của nó sẽ được tính bằng :

$inside(N^j)=lg(P(N^j \rightarrow N^r N^s)) + inside(N^r) + inside(N^s)$

Hình 3 Xác suất inside

inside của một nút ở đây mang ý nghĩa là xác suất của nút đó, trong trường hợp nút này là nút gốc S, thì inside chính là xác suất của cây. [1]

3.1.2 Xác suất outside.

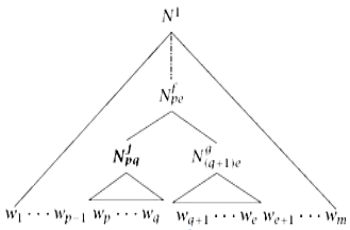
Giả sử ta có nút D và có hai luật $A \rightarrow B\ D$ và $C \rightarrow D\ E$ thì outside của D sẽ được tính bằng:

$P1 = lgP((A \rightarrow B\ D)) + inside(B) + outside(A)$

$P2 = lg(P(C \rightarrow D\ E)) + inside(E) + outside(C)$

$Outside(D) = max(P1,P2);$

Outside của một nút ở đây mang ý nghĩa tượng trưng cho độ liên kết của nút với nút gốc, hay nói một cách khác nó là xác suất của một nút về việc từ nút đó có bao nhiêu khả năng tìm được nút gốc. [2]



Hình 4 Xác suất outside

3.2 Giải thuật A-star.

Bộ phân tích sẽ tạo ra hai tập : agenda và chart. Trong đó agenda là tập nút đang xếp hàng chờ được xem xét, còn chart là tập các nút đã xét qua. Bộ phân tích cú pháp sẽ lấy ra nút có độ ưu tiên cao nhất, kết hợp với các phần tử trong chart, tạo ra một tập các nút mới, các nút này sẽ được thêm vào agenda để chờ xử lí tiếp. Bộ phân tích sẽ dừng lại khi tìm được S[1,n+1] là đáp áp cuối cùng hoặc không còn nút để xét. Nếu kết thúc giải thuật, tìm được đáp án cuối cùng thì quá trình phân tích thành công, ngược lại quá trình phân tích thất bại.[3]

Vấn đề lớn nhất ở đây chính là hàm mục tiêu để chọn ra phân tử có độ ưu tiên cao nhất. Hàm mục tiêu trong A-star bao giờ cũng có hai thành phần là chi phí và ước lượng. Chi phí là quãng đường đã xét duyệt qua, ở đây ta có thể dùng inside như một hàm chi phí. Còn về ước lượng quãng đường đi đến đích, bản thân outside đặc trưng cho khả năng tìm được đường đến nút gốc nên outside là một sự lựa chọn tuyệt vời. Vậy công thức hàm mục tiêu của chúng ta sẽ như dưới đây:

f(nút) = inside(nút) + outside(nút).

3.3 Kết quả và đánh giá.

Xác suất của các luật trong bộ phân tích cú pháp được tính toán dựa vào việc thống kê từ tập huấn luyện viettreebank gồm 2000 câu đã được phân tích đúng 100%.

Tập dữ liệu test	Phần trăm phân tích được	Thời gian
630 câu bất kì	92%	20mins

Tập dữ liệu test	Phần trăm phân tích chính xác	Thời gian
500 câu bất kì	60-70%	15mins

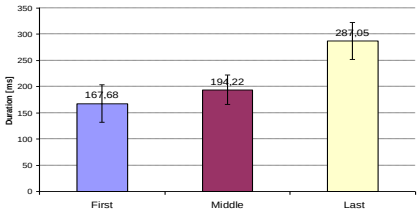
4. MÔ HÌNH HÓA TRƯỜNG ĐỘ ÂM TIẾT TIẾNG VIỆT.

Trong tất cả các hệ thống tổng hợp tiếng nói, để có thể đạt được độ tự nhiên cao cho tiếng nói tổng hợp, một vấn đề cần phải xử lý đó là dự đoán được trường độ cho các âm tiết. Một trong những phương pháp truyền thống trong việc sử mô hình hóa trường độ đó là sử dụng các tập luật, nhưng việc xây dựng và chọn lựa được các luật là một công việc rất khó và đạt độ chính xác không cao.

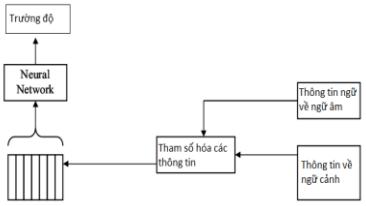
Cùng với sự phát triển của trí tuệ nhân tạo, và học máy đã có nhiều phương pháp huấn luyện đạt được độ chính xác cao trong việc dự đoán trường độ như sử dụng cây quyết định hoặc mạng Neuron. Trong đó mạng Neuron là phương pháp đạt được độ chính xác cao nhất đối với việc đưa ra trường độ của âm tiết.

4.1 Các yếu tố ảnh hưởng đến trường độ âm tiết

Có nhiều yếu tố ảnh hưởng đến trường độ của âm tiết trong tiếng Việt, các yếu tố này có thể phân thành các nhóm chính. Ngữ âm, ngữ cảnh, và vị trí của âm tiết. Ví dụ đối với cùng một âm tiết thì nếu như âm tiết đó đứng ở vị trí cuối câu hoặc cuối từ thì trường độ của âm tiết đó dài hơn hẳn so với các thể hiện của âm tiết đó ở vị trí khác.[8]



Hình 2 Ảnh hưởng của vị trí đến trường độ của âm tiết

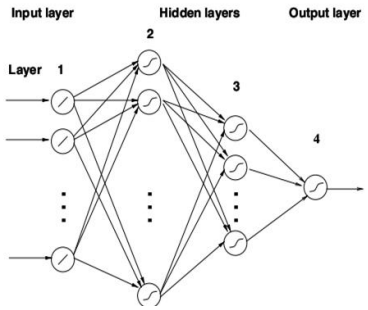


Hình 3 Mô hình hóa trường độ

Bộ phân tích cú pháp sẽ xử lý đối với các văn bản đầu vào để đưa ra thông tin về ngữ cảnh của âm tiết như là các âm tiết liền trước, liền sau, vị trí của âm tiết trong câu… Về mặt ngữ âm, thì một âm tiết tiếng Việt thuộc một trong tám loại V, VC, CV, CVC, VV, VVC, CVV, CVVC. Trong đó V là nguyên âm và C là phụ âm (Vowel and Consonant)[11]. Dựa vào đó, các âm tiết sẽ được phân tích để đưa ra các thông tin về mặt ngữ âm như các thành phần cấu tạo nên âm tiết, số lượng thành phần của âm tiết, thanh điệu. Các thông tin này được tham số hóa về trong khoảng (0,1) để làm đầu vào cho mạng neuron. Giá trị đầu ra của mạng Neuron chính là giá trị về trường độ của âm tiết.

4.2 Mạng Neuron

Mạng Neuron sử dụng trong quá trình huấn luyện là mô hình mạng dẫn tiến nhiều lớp. Sử dụng giải thuật học lan truyền ngược (Backpropagation)[12]. Với một tầng đầu vào, hai tầng ẩn và một đầu ra. Kiến trúc của mạng Neuron được lựa chọn theo phương pháp thử sai.



Hình 4. Mạng Neuron

4.3 Kết quả

Trường độ của âm tiết được dự đoán đạt độ chính xác 85% với dữ liệu đã được huấn luyện. Và độ chính xác 82% với các âm tiết nằm ngoài tập huấn luyện. Kết quả này đạt độ chính xác tương đối cao so với các hệ thống hệ thống sử dụng CART 77% [8].

5 MÔ HÌNH HÓA CAO ĐỘ

Trong các hệ thống tổng hợp tiếng nói, việc sinh ra đường F0 là một vấn đề thiết yếu để có thể thu được những âm thanh

tổng hợp tự nhiên. Hiện nay đã có một số mô hình sinh ra đường tần số cơ bản (F0) của tiếng Việt. Các mô hình này đã đạt được những kết quả đáng chú ý, nhưng đều mới là mô hình ngữ điệu cho câu khẳng định.

Trong nghiên cứu này, chúng tôi xem xét vai trò của hai yếu tố trong ngữ điệu của câu hỏi là ngữ điệu của toàn câu và ngữ điệu của phần cuối câu. Đầu tiên một bộ dữ liệu được xây dựng, sau đó chúng tôi thực hiện biến đổi F0 của các câu này theo hai yếu tố và thực hiện một bài kiểm tra cảm thụ để đánh giá được vai trò của chúng.

5.1 Khác nhau giữa ngữ điệu câu hỏi và câu khẳng định

Theo [8], ngữ điệu của câu hỏi có xu hướng đi lên cao ở cuối câu mà không bị ảnh hưởng bởi âm tiết cuối là mang thanh điệu nào. [8], [24] và [25] đưa ra kết luận là câu hỏi được nói với một âm vực cao hơn câu hỏi. Do đó, trong nghiên cứu này, chúng tôi sẽ xem xét ảnh hưởng của hai yếu tố này trong ngữ điệu của câu hỏi, nhằm đề xuất ra được một mô hình ngữ điệu cho câu hỏi.

5.2 Bộ dữ liệu

Bộ dữ liệu bao gồm 25 câu khẳng định, với nội dung trích ra từ các đoạn hội thoại trong cuộc sống hằng ngày. Mục đích của việc lựa chọn các câu đối thoại là có thể dễ dàng biến đổi sang các câu hỏi nghi vấn với nội dung giống hệt nhưng với ý định hỏi để xác nhận.

5.3 Thí nghiệm thay đổi F0 để đạt ngữ điệu câu hỏi

5.3.1 Ảnh hưởng của ngữ điệu toàn câu

Phương pháp

Bài test thứ nhất nhằm xác định ảnh hưởng của ngữ điệu toàn câu với trường hợp của câu hỏi. Chúng tôi thực hiện việc nâng cao ngữ điệu của 25 câu đã có lên lần lượt 2 mức là 10% F0 trung bình của cả câu và 20% F0 trung bình của cả câu. Sau đó 50 câu này được trộn với 25 câu gốc tạo ra một bộ dữ liệu test gồm 75 câu. Bộ dữ liệu này sau đó được 9 nam và 9 nữ nghe và đánh giá xem những câu này có giống câu hỏi không, và mức độ tự tin của họ khi đưa ra những đánh giá đó.

Kết quả

Mức độ tăng	Tỉ lệ chọn câu hỏi	Độ lệch chuẩn của tỉ lệ chọn	Mức độ tin tưởng
10%	50.22%	23.93%	75.82%
20%	76.00%	17.66%	82.28%

Bảng 2 Thống kê kết quả đối với 9 nam

Mức độ tăng	Tỉ lệ chọn câu hỏi	Độ lệch chuẩn của tỉ lệ chọn	Mức độ tin tưởng
10%	35.56%	20.14%	77.59%
20%	73.33%	12.81%	77.99%

Bảng 3 Thống kê kết quả đối với 9 nữ

Mức độ tăng	Tỉ lệ chọn câu hỏi	Độ lệch chuẩn của tỉ lệ chọn	Mức độ tin tưởng
10%	42.89%	22.74%	76.70%
20%	74.66%	15.02%	80.13%

Bảng 4 Thống kê kết quả chung

5.3.2 Ảnh hưởng của ngữ điệu cuối câu

Phương pháp

Bài test thứ hai cũng được thực hiện tương tự bài test thứ nhất. Điểm khác biệt duy nhất là 25 câu đã có được nâng cao ngữ điệu của âm tiết cuối lên lần lượt 2 mức là 10% F0 trung bình cả câu và 20% F0 trung bình cả câu. Công việc này cũng được thực hiện hoàn toàn tự động bằng phần mềm Praat.

Kết quả

Mức độ tăng	Tỉ lệ chọn câu hỏi	Độ lệch chuẩn của tỉ lệ chọn	Mức độ tin tưởng
10%	55.11%	22.34%	78.61%
20%	80.44%	16.78%	85.55%

Bảng 5 Thống kê kết quả đối với 9 nam

Mức độ tăng	Tỉ lệ chọn câu hỏi	Độ lệch chuẩn của tỉ lệ chọn	Mức độ tin tưởng
10%	43.11%	15.59%	75.32%
20%	72.89%	13.08%	78.44%

Bảng 6 Thống kê kết quả đối với 9 nữ

Mức độ tăng	Tỉ lệ chọn câu hỏi	Độ lệch chuẩn của tỉ lệ chọn	Mức độ tin tưởng
10%	49.11%	19.68%	76.97%
20%	76.67%	15.10%	82.00%

Bảng 7 Thống kê kết quả chung

5.4 Kết luận

Dựa vào kết quả của hai bài test ta có thể thấy vai trò của hai yếu tố trong ngữ điệu của câu hỏi là gần tương đương nhau. Ngữ điệu của cuối câu khi được nâng lên cho kết quả cao hơn một chút so với khi ngữ điệu của cả câu được nâng lên. Khi mức nâng là 20% với cả hai trường hợp thì tỉ lệ chọn là câu hỏi đã lên đến xấp xỉ 75%, đây là kết quả khá tốt trong trường hợp câu hỏi không có từ để hỏi. Do vậy, khi áp dụng vào hệ tổng hợp tiếng nói, ta cần kết hợp giữa cả hai yếu tố này sao cho hợp lý để đạt kết quả cao nhất.

6. TỔNG HỢP MỨC THẤP

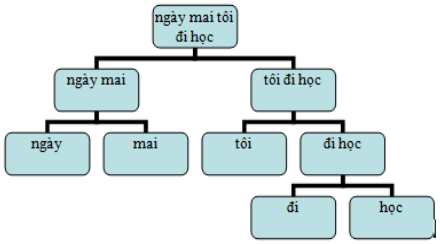
Tổng hợp mức thấp là quá trình lựa chọn, tìm kiếm đơn vị âm trong cơ sở dữ liệu đơn vị âm và ghép nối chúng để tạo nên tiếng nói cần tổng hợp. Các loại đơn vị âm có thể dùng để ghép nối với chiều dài tăng dần là âm vị, bán âm tiết, âm đầu/vần, âm tiết, cụm từ. Theo [8], trong tiếng Việt, loại đơn vị âm có thể dùng để tổng hợp cho chất lượng tốt là bán âm tiết, âm tiết, cụm từ. Quan điểm được thừa nhận rộng rãi hiện nay là đơn vị âm càng lớn thì chất lượng tiếng nói tổng hợp càng cao do giảm thiểu được số điểm ghép nối tín hiệu, tuy nhiên đổi lại là sự tăng lên về kích thước cơ sở dữ liệu. Với mục tiêu là nâng cao chất lượng tiếng nói tổng hợp, giải pháp là sử dụng thuật toán lựa chọn đơn vị không đồng nhất kết hợp với phương pháp TD-PSOLA để điều khiển các tham số ngữ điệu tiếng nói.

6.1. Lựa chọn đơn vị

Các loại đơn vị được sử dụng với mức ưu tiên giảm dần là cụm từ, âm tiết, bán âm tiết. Câu cần tổng hợp sẽ được chia ra thành các cụm từ theo các mức khác nhau nhờ quá trình phân tích cú pháp. Ví dụ như cho câu "Ngày mai tôi đi học".

Quá trình tìm kiếm sẽ được bắt đầu từ gốc, sau đó đi xuống các nhánh. Việc tìm kiếm sẽ dừng lại ở mức cao nhất có thể ngay khi tìm thấy cụm từ hoặc đi tới mức lá là các âm tiết. Cách thức phân chia để tìm kiếm này làm tăng xác suất tìm thấy của những cụm từ có độ dài lớn hơn một âm tiết hơn là việc chọn ngẫu nhiên cụm từ theo một độ dài xác định nào đó để tìm kiếm. Đây là ý tưởng chủ đạo trong thuật toán lựa chọn đơn vị không đồng nhất.

Trong trường hợp không tìm thấy ứng viên nào ở mức lá, âm tiết còn lại sẽ được tổng hợp ở mức bán âm tiết. Theo [8], việc tổng hợp ở mức bán âm tiết có thể tổng hợp được hầu hết các âm tiết trong tiếng Việt.



Hình 8 Cây phân tích cú pháp

6.2. Hàm chi phí

Kết quả của quá trình tìm kiếm đơn vị sẽ là tập các mẫu của các đơn vị âm tìm thấy. Một đơn vị âm có thể có nhiều mẫu. Nhiệm vụ là cần chọn ra mẫu tốt nhất trong đó để ghép nối. Theo [8], mẫu được chọn sẽ dựa theo tiêu chí có sự khác biệt nhỏ nhất về ngữ điệu với đơn vị âm đích mong muốn. Sự sai khác này được lượng hóa thành hàm chi phí. Hàm chi phí là tổng của tất cả các độ méo bao gồm hai loại độ méo chính:

– Độ méo của đơn vị âm thể hiện bằng sự khác nhau giữa đơn vị âm được lựa chọn với đơn vị âm cần tổng hợp. Đây gọi là chi phí đích.

– Độ méo về sự liên tục được thể hiện bằng khoảng cách giữa đơn vị âm được chọn so với đơn vị âm trước đó. Đây gọi là chi phí ghép nối.

Thuật toán Viterbi được sử dụng để chọn ra các đơn vị âm có hàm chi phí nhỏ nhất. Sau đó, các đơn vị âm được ghép nối và làm trơn bằng thuật toán TD-PSOLA.

6.3 Kết quả tổng hợp

Chương trình đã tổng hợp một đoạn văn bản với số lượng 11 câu. Kết quả tiếng nói tổng hợp ban đầu tương đối dễ nghe, tuy nhiên về mức độ tự nhiên và ngữ điệu còn hạn chế. Việc này cần được cải tiến bằng cách thay đổi các tham số trong hàm chi phí, sử dụng thuật toán làm mượt các điểm ghép nối.

7. KẾT LUẬN

Một mô hình hệ thống tổng hợp tiếng nói với năm module được đề xuất nhằm cải thiện chất lượng tiếng nói tổng hợp. Việc chuẩn hóa văn bản đầu vào giúp tổng hợp được nhiều trường hợp viết tắt và nhập nhằng trong tiếng Việt. Phân tích cú pháp và lựa chọn đơn vị ghép nối đảm bảo độ dễ nghe của âm thanh nhờ giảm thiểu được chi phí ghép nối. Và nhờ vào phân tích ngữ điệu (trường độ và cao độ) âm thanh tổng hợp có được độ tự nhiên cao.

Trong thời gian tới, nhóm chúng tối sẽ tập trung vào cải tiến các module, ráp nối và hoàn thiện để hệ thống hoạt động ổn định, đạt chất lượng cao.

8. LỜI TRI ÂN

Chúng tôi xin gửi lời cảm ơn chân thành TS. Trần Đỗ Đạt tại Trung tâm nghiên cứu Mica và ThS. Nguyễn Thị Thu Trang – Bộ môn Công nghệ phần mềm – Viện Công nghệ thông tin và truyền thông đã hết lòng hướng dẫn và chỉ bảo chúng tôi trong suốt quá trình nghiên cứu.

Chúng tôi cũng bày tỏ lòng biết ơn trung tâm nghiên cứu Mica đã tạo điều kiện về cơ sở vật chất cho chúng tôi trong quá trình học tập và nghiên cứu.

Cuối cùng, chúng tôi xin cảm ơn các thầy cô giáo trong Viện Công nghệ thông tin và truyền thông đã giảng dạy cho chúng tôi các kiến thức hữu ích trong suốt những năm vừa qua.

9. TÀI LIỆU THAM KHẢO

[1] Fei Xia, "Inside-Outside algorithm", LING 572.

[2] Michael Collins, "*Head-Driven Statistical Models for Natural Language Parsing*", MIT Computer Science and Artificial Intelligence Laboratory.

[3] Dan Klein and Christopher D. Manning. 2003. "*A* parsing: Fast exact Viterbi parse selection. In Proceedings of the Human Language Technology Conference and the North American Association for Computational Linguistics*"(HLT-NAACL).

[4] Adam Pauls and Dan Klein, "*K-Best A* Parsing*", Computer Science Division University of California, Berkeley.

[5] Hoàng Anh Việt, "*Phân tích cú pháp tiếng việt sử dụng mô hình xác suất PCFG*", đồ án tốt nghiệp đại học năm 2006.

[6] Phạm Thị Nhung, "*Phân tích cú pháp tiếng việt sử dụng beam search*", đồ án tốt nghiệp đại học năm 2009.

[7] Đỗ Bá Lâm, Lê Thanh Hương, "*Implementing a Vietnamese syntactic parser using HPSG*", Khoa Công nghệ thông tin, trường Đại học Bách khoa Hà Nội.

[8] Trần Đỗ Đạt, "*Synthèse de la parole a partir du texte en langue Vietnamienne*", Thèse en cotutelle international MICA, Hanoi, 2007.

[9] Minghui Dong, Kim-Teng Lua, Haizhou Li, "*A Unit Selection-based Speech Synthesis Approach for Mandarin Chinese*", Institute for Infocomm Research.

[10] Vũ Hải Quân, Cao Xuân Nam, *"Tổng hợp tiếng nói tiếng Việt, theo phương pháp ghép nối cụm từ".*

[11] Trần Ngọc Dụng, *"Cẩm nang văn phạm tiếng Việt"*, 2010.

[12] Christopher M. Bishop, *"Neural network for pattern regconition"*.

[13] Nguyễn Thị Thu Trang, Phạm Thị Thanh, Trần Đỗ Đạt, *"A method for Vietnamese Text Normalization to improve the quality of speech synthesis".*

[14] Taylor Paul. *"Text-To-Speech Synthesis. s.l. : Cambridge University Press, 2009"*

[15] Language Technologies Institute Carnegie Mellon University, *"Non – Standard Word and Homograph Resolution for Asian Language Text Analysis"*

[16] Richard Sproat, Alan Black, Stanley Chen, Shankar Kumar, Mari Osten-dorf, Christopher Richards, *"Normalization of Non-StandardWords. Computer Speech and Language, Volume 15, Issue 3. July 2001".*

[17] Stanley F. Chen, Joshua Goodman, *"An Empirical Study of Smoothing Techniques for Language Modeling"*

[18] K. Sreenivasa Rao, B. Yegnanarayana, "*Modeling durations of syllables using neural network",* ScienceDirect, 2006.

[19] Martti Vainio, "*Artificial Neural Network Based Prosody Models for Finnish",* University of Helsinki, Department of Phonetics.

[20] Marcello Balestri, Alberto Pacchiotti, Silvia Quazza, Pier Luigi Salza, Stefano Sandri, "*Choose the best to modify the least: a new generation concatenative synthesis system*", CSELT - Centro Studi e Laboratori Telecomunicazioni S.p.A., Torino, Italy.

[21] Min Chu,Hu Peng, Hong-yunYang, Eric Chang, "*Selecting non-uniform units from a very large corpus for cancatenative speech synthesizer*", Microsoft Research China, Beijing.

[22] Paul Taylor, "*Text-to-SpeechSynthesis",* University of Cambridge, Cambridge University Press.

[23] Mark Tatham, Katherine Morton, "*Development in Speech Synthesis*", Wiley, 2005.

[24] Đỗ T.D., Trần T.H., et Boulakia G., *"Intonation in vietnamese"*, Intonation systems: A survey of 22 languages, Hirst & Di Cristo (ed.), Cambridge U.P, 1998.

[25] Nguyễn T.T.H, *"Contribution à l'étude de la prosodie du vietnamien. Variations de l'intonation dans les modalités: assertive, interrogative et impérative"*, Thèse 2004, Doctorat de Linguistique Théorique, Formelle et Automatique.