Tách từ tiếng Việt

Lê Thanh Hương
Bộ môn Hệ thống Thông tin
Viện CNTT &TT – Trường ĐHBKHN
Email: huonglt-fit@mail.hut.edu.vn

Tách từ

- Mục đích: xác định ranh giới của các từ trong câu.
- Là bước xử lý quan trọng đối với các hệ thống XLNNTN, đặc biệt là đối với các ngôn ngữ đơn lập, ví dụ: âm tiết Trung Quốc, âm tiết Nhật, âm tiết Thái, và tiếng Việt.
- Với các ngôn ngữ đơn lập, một từ có thể có một hoặc nhiều âm tiết.
  - Vấn đề của bài toán tách từ là khử được sự nhập nhằng trong ranh giới từ.

Từ vựng

- tiếng Việt là ngôn ngữ không biến hình
- Từ điển từ tiếng Việt (Vietlex): >40.000 từ, trong đó:
  - 81.55% âm tiết là từ : từ đơn
  - 15.69% các từ trong từ điển là từ đơn
  - 70.72% từ ghép có 2 âm tiết
  - 13.59% từ ghép ≥ 3 âm tiết
  - 1.04% từ ghép ≥ 4 âm tiết

Từ vựng

Độ dài	#	%
1	6,303	15.69
2	28,416	70.72
3	2,259	5.62
4	2,784	6.93
5	419	1.04
Tổng	40,181	100

Bảng 1. Độ dài của từ tính theo âm tiết

Qui tắc cấu tạo từ tiếng Việt

- Từ đơn: *dùng một âm tiết* làm một từ.
  - Ví dụ: *tôi, bác, người, cây, hoa, đi, chạy, vì, đã, à, nhỉ, nhé...*
- Từ ghép: *tổ hợp* (ghép) các âm tiết lại, giữa các âm tiết đó có quan hệ về nghĩa với nhau.
  - *Từ ghép đẳng lập*. các thành tố cấu tạo có quan hệ bình đẳng với nhau về nghĩa.
    - Ví dụ: *chợ búa, bếp núc*
  - *Từ ghép chính phụ*. các thành tố cấu tạo này phụ thuộc vào thành tố cấu tạo kia. Thành tố phụ có vai trò phân loại, chuyên biệt hoá và sắc thái hoá cho thành tố chính.
    - Ví dụ: *tàu hoả, đường sắt, xấu bụng, tốt mã, ngay đơ, thẳng tắp, sưng vù...*

Qui tắc cấu tạo từ tiếng Việt

- Từ láy: các yếu tố cấu tạo có thành phần ngữ âm được lặp lại; nhưng vừa lặp vừa biến đổi. Một từ được lặp lại cũng cho ta từ láy.
- Biến thể của từ: được coi là *dạng lâm thời biến động* hoặc *dạng "lời nói"* của từ.
  - Rút gọn một từ dài thành từ ngắn hơn
    - ki-lô-gam → ki lô/ kí lô
  - Lâm thời phá vỡ cấu trúc của từ, phân bố lại yếu tố tạo từ với những yếu tố khác ngoài từ chen vào. Ví dụ:
    - khổ sở → lo khổ lo sở
    - ngặt nghèo → cười ngặt cười nghẽo
    - danh lợi + ham chuộng → ham danh chuộng lợi

Qui tắc cấu tạo từ tiếng Việt

- Các diễn tả gồm nhiều từ (vd, "bởi vì") cũng được coi là 1 từ
- Tên riêng: tên người và vị trí được coi là 1 đơn vị từ vựng
- Các mẫu thường xuyên: số, thời gian

Các hướng tiếp cận

- Tiếp cận dựa trên từ điển
- Tiếp cận theo phương pháp thống kê
- Kết hợp hai phương pháp trên.

Các phương pháp

- So khớp từ dài nhất (Longest Matching)
- Học dựa trên sự cải biến (Transformation-based Learning – TBL)
- Chuyển đổi trạng thái trọng số hữu hạn (Weighted Finite State Transducer – WFST)
- Độ hỗn loạn cực đại (Maximum Entropy – ME)
- Học máy sử dụng mô hình Markov ẩn (Hidden Markov Models- HMM)
- Học máy sử dụng vectơ hỗ trợ (Support Vector Machines)
- Kết hợp một số phương pháp trên

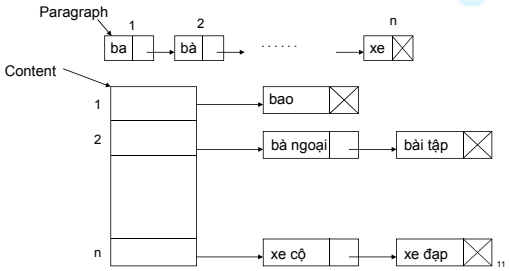
Tiếp cận dựa trên từ điển

<Lê Thanh Hương, Phân tích cú pháp tiếng Việt, Luận văn cao học, 1999>

- Xây dựng từ điển
  - Mỗi mục từ lưu thông tin về từ, từ loại, nghĩa loại
  - Tổ chức sao cho tốn ít bộ nhớ và thuận tiện trong việc tìm kiếm
- Mã hóa từ điển: Từ loại và nghĩa loại kiểu byte được lưu dưới dạng một ký tự.
- VD: danh từ -112 – p, <loại từ> - 115 – s

Tiếp cận dựa trên từ điển

- Phân trang theo hai chữ cái đầu của từ, sắp tăng. Với mỗi trang, các từ lại được sắp theo vần ABC.



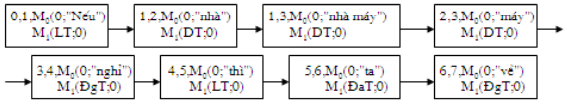
Tìm từ trong từ điển

- Độ dài tối đa của từ? 3? 4? 5?
- Vấn đề: không xử lý được các tổ hợp từ cố định, vd "ông chẳng bà chuộc"
- ➢ *Đưa ra tất cả các từ ghép có trong từ điển trùng với phần đầu của xâu vào*

Tìm từ trong từ điển

	Nếu	nhà	máy	nghỉ	thì	ta	về
Vị trí từ:	0	1	2	3	4	5	6	7

- Ta có bảng sau:



- Ký hiệu:
  - <liên từ> - LT <danh từ> - DT
  - <động từ> - ĐgT <đại từ> - ĐaT

13

Phân giải nhập nhằng

- Lấy tất cả các cách phân tích, nếu phân tích cú pháp cho ra cây đúng thì đó là cách phân tích đúng.

14

Cách tiếp cận lai

<Phuong Le-Hong et al., A hybrid approach to word segmentation of Vietnamese texts, Proceedings of the 2nd International Conference on Language and Automat Theory and Applications, LATA 2008, Tarragona, Spain, 2008.>

- Kết hợp phân tích automat hữu hạn + biểu thức chính quy + so khớp từ dài nhất + thống kê (để giải quyết nhập nhằng)

15

Biểu thức chính qui

- là một khuôn mẫu được so sánh với một chuỗi
- Các ký tự đặc biệt:
  - * - bất cứ chuỗi ký tự nào, kể cả không có gì
  - x – ít nhất 1 ký tự
  - + - chuỗi trong ngoặc xuất hiện ít nhất 1 lần
- Ví dụ:
  - Email: x@x(.x)+
  - dir *.txt
  - '*John' -> 'John', 'Ajohn', "Decker John"
- Biểu thức chính quy được sử dụng đặc biệt nhiều trong:
  * Phân tích cú pháp
  * Xác nhận tính hợp lệ của dữ liệu
  * Xử lý chuỗi
  * Tách dữ liệu và tạo báo cáo

16

Automat hữu hạn

- Lớp ngôn ngữ chính qui, được đoán nhận bởi máy ảo, gọi tên là automat hữu hạn.
  - Automat hữu hạn đơn định (Deterministic Finite Automat a– DFA
  - Automat hữu hạn không đơn định (Nondeterministic Finite Automat a– NFA)
  - Automat hữu hạn không đơn định, chấp nhận phép truyền rỗng (ε-NFA)

17

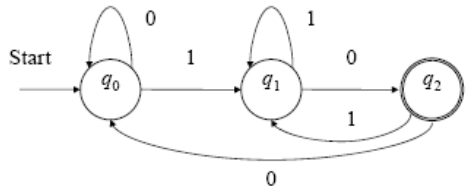
Giới thiệu phi hình thức về automat hữu hạn

- Một bài toán trong automat là nhận diện chuỗi *w* có thuộc về ngôn ngữ *L hay không.*
- Chuỗi nhập được xử lý tuần tự từng ký hiệu một từ trái sang phải.
- Trong quá trình thực thi, automat cần phải nhớ thông tin đã qua xử lý.

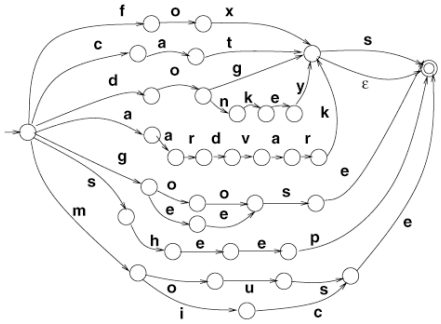
18

Ví dụ về automat hữu hạn

L = {w ∈ {0, 1}* | w kết thúc bằng chuỗi con 10}.



Automat hữu hạn cho các từ tiếng Anh

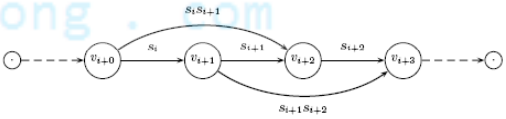


Cách tách từ đơn giản

- Phát hiện các mẫu thông thường như tên riêng, chữ viết tắt, số, ngày tháng, địa chỉ email, URL,… sử dụng biểu thức chính qui
- Hệ thống chọn chuỗi âm tiết dài nhất từ vị trí hiện tại và có trong từ điển, chọn cách tách có ít từ nhất
  - Hạn chế: có thể đưa ra cách phân tích không đúng.
  - Giải quyết: liệt kê tất, có 1 chiến lược để chọn cách tách tốt nhất.

Lựa chọn cách tách từ

- Biểu diễn đoạn bằng chuỗi các âm tiết $s_1 s_2 \ldots s_n$
- Trường hợp nhập nhằng thường xuyên nhất là 3 từ liền nhau $s_1 s_2 s_3$ trong đó $s_1 s_2$ và $s_2 s_3$ đều là từ.



- Biểu diễn 1 đoạn bằng đồ thị có hướng tuyến tính $G = (V,E)$, $V = \{v_0, v_1, \ldots, v_n, v_{n+1}\}$
- Nếu các âm tiết $s_{i+1}, s_{i+2}, \ldots, s_j$ tạo thành 1 từ -> trong G có cạnh (v_i, v_j)
- Các cách tách từ = các đường đi ngắn nhất từ v_0 đến v_{n+1}

Thuật toán

Thuật toán 1. Xây dựng đồ thị cho chuỗi $s_1 s_2 \ldots s_n$

```
1: V ← ∅;
2: for i = 0 to n + 1 do
3:     V ← V ∪ {v_i};
4: end for
5: for i = 0 to n do
6:     for j = i to n do
7:         if (accept(A_W, s_i · · · s_j)) then
8:             E ← E ∪ {(v_i, v_{j+1})};
9:         end if
10:    end for
11: end for
12: return G = (V,E);
```

accept(*A, s*): automat A nhận xâu vào s

Phân giải nhập nhằng

- Xác suất xâu s:

$$P(s) = \prod_{i=1}^{m} P(w_i|w_1^{i-1}) \approx \prod_{i=1}^{m} P(w_i|w_{i-n+1}^{i-1})$$

- $P(w_i|w_1^{i-1})$: xác suất w_i khi có i-1 âm tiết trước đó
- n = 2: bigram; n = 3: trigram

Phân giải nhập nhằng

- Khi n = 2, tính giá trị $P(w_i|w_{i-1})$ lớn nhất maximum likelihood (ML)

$$P_{ML}(w_i|w_{i-1}) = \frac{P(w_{i-1}w_i)}{P(w_{i-1})} = \frac{c(w_{i-1}w_i)/N}{c(w_{i-1})/N} = \frac{c(w_{i-1}w_i)}{c(w_{i-1})}$$

- c(s): số lần xâu s xuất hiện; N: tổng số từ trong tập luyện
- Khi dữ liệu luyện nhỏ hơn kích cỡ toàn bộ tập dữ liệu → P ~ 0
- Sử dụng kỹ thuật làm trơn

25

Kỹ thuật làm trơn

$$\widehat{P}(w_i|w_{i-1}) = \lambda_1 P_{ML}(w_i|w_{i-1}) + \lambda_2 P_{ML}(w_i)$$

với $\lambda_1 + \lambda_2 = 1$ và $\lambda_1, \lambda_2 \geq 0$

$P_{ML}(w_i) = c(w_i)/N$

- Với tập thử nghiệm T = $\{s_1, s_2, \ldots, s_n\}$, xác suất P(T) của tập thử: $P(T) = \prod_{i=1}^{n} P(s_i)$
- Entropy của văn bản:

$$H_p(T) = \frac{-\log_2 P(T)}{N_T} = -\frac{1}{N_T}\sum_{i=1}^{n} \log_2 P(s_i)$$

với N_T: số từ trong T

- Entropy tỉ lệ nghịch với xác suất trung bình của 1 cách tách từ cho các câu trong văn bản thử nghiệm.

26

Xác định giá trị λ_1, λ_2

- Từ tập dữ liệu mẫu, định nghĩa $C(w_{i-1}, w_i)$ là số lần (w_{i-1}, w_i) xuất hiện trong tập mẫu. Ta cần chọn λ_1 λ_2 để làm cực đại giá trị

$$L(\lambda_1, \lambda_2) = \sum_{w_{i-1}, w_i} C(w_{i-1}, w_i) \log_2 \widehat{P}(w_i|w_{i-1})$$

với $\lambda_1 + \lambda_2 = 1$ và $\lambda_1, \lambda_2 \geq 0$

Thuật toán

Thuật toán 2. Xác định giá trị λ

$\lambda_1 \leftarrow 0.5, \lambda_2 \leftarrow 0.5;$
$\epsilon \leftarrow 0.01;$
repeat
  $\widehat{\lambda}_1 \leftarrow \lambda_1, \quad \widehat{\lambda}_2 \leftarrow \lambda_2;$
  $c_1 \leftarrow \sum_{w_{i-1}, w_i} \frac{C(w_{i-1}, w_i)\lambda_1 P_{ML}(w_i|w_{i-1})}{\lambda_1 P_{ML}(w_i|w_{i-1}) + \lambda_2 P_{ML}(w_i)};$
  $c_2 \leftarrow \sum_{w_{i-1}, w_i} \frac{C(w_{i-1}, w_i)\lambda_2 P_{ML}(w_i)}{\lambda_1 P_{ML}(w_i|w_{i-1}) + \lambda_2 P_{ML}(w_i)};$
  $\lambda_1 \leftarrow \frac{c_1}{c_1 + c_2}, \quad \lambda_2 \leftarrow 1 - \widehat{\lambda}_1;$
  $\widehat{\epsilon} \leftarrow \sqrt{(\widehat{\lambda}_1 - \lambda_1)^2 + (\widehat{\lambda}_2 - \lambda_2)^2};$
until $(\widehat{\epsilon} \leq \epsilon);$
return $\lambda_1, \lambda_2;$

28

Kết quả

- Sử dụng tập dữ liệu gồm 1264 bài trong báo Tuổi trẻ, có 507,358 từ
- Lấy $\varepsilon = 0.03$, các giá trị λ hội tụ sau 4 vòng lặp

Step	λ_1	λ_2	ϵ
0	0.500	0.500	1.000
1	0.853	0.147	0.499
2	0.952	0.048	0.139
3	0.981	0.019	0.041
4	0.991	0.009	0.015

- Độ chính xác = số từ hệ thống xác định đúng/tổng số từ hệ thống xác định = 95%

29