ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Tách từ tiếng Việt

Viện Công nghệ Thông tin và Truyền thông

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

- Tiếng Việt là ngôn ngữ không biến hình
- Từ điển từ tiếng Việt (Vietlex): >40.000 từ

Độ dài	# từ	%
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 nghẽo → 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 dựa trên học máy
- Kết hợp hai phương pháp trên.

Tách từ dựa trên từ điển

- Thuật toán so khớp từ dài nhất
- Yêu cầu:
  - Từ điển
  - Chuỗi đầu vào đã tách các dấu câu và âm tiết
- Tư tưởng: thuật toán tham lam
  - Đi từ trái sang phải hoặc từ phải sang trái, lấy các từ dài nhất có thể, dừng lại khi duyệt hết
  - Độ phức tạp tính toán: *O(n . V)*
    - *n: Số âm tiết trong chuỗi*
    - *V: Số từ trong từ điển*

Tách từ dựa trên từ điển

- Thuật toán so khớp từ dài nhất

- **BẮT ĐẦU** — *khởi tạo*
- (1) Cho chuỗi đầu vào [*w_0 w_1 … w_n-1*]
- (2) *words* ← *[]*
- (3) *s* ← *0*

- (4) *e* ← *n* — *lặp*
- (5) Khi [*w_s … w_e*] **chưa là một từ**: *e* ← *e – 1*
- (6) *words* ← *words* + *[w_s … w_e]*
- (7) *s* ← *e* + *1*
- (8) Nếu *e* < *n*: Quay lại bước (4)

- (9) Lấy ra chuỗi đã tách từ *words* — *kết thúc*
- **KẾT THÚC**
