

Mô hình tách từ, gán nhãn từ loại và hướng tiếp cận tích hợp cho tiếng Việt : Luận văn ThS / Trần Thị Oanh ; Nghd. : PGS.TS. Hà Quang Thụy . - H. : ĐHCN, 2008 . - 48 tr. + CD-ROM

Tóm tắt: Giới thiệu hai bài toán cơ bản trong xử lý ngôn ngữ tự nhiên là bài toán tách từ và bài toán gán nhãn từ loại tiếng Việt, các hướng tiếp cận cho mỗi bài toán và tình hình nghiên cứu chung ở Việt Nam cũng như trên thế giới. Nghiên cứu và đề xuất mô hình cho bài toán tách từ tiếng Việt tận dụng thông tin tri thức từ nhiều nguồn khác nhau nhằm tăng độ chính xác của bộ tách từ. Nghiên cứu và đề xuất mô hình gán nhãn từ loại tiếng Việt, xây dựng một mô hình gán nhãn từ loại hiệu quả như thiết kết corpus gán nhãn từ loại, thiết kế các tập đặc trưng khác nhau để tìm ra các đặc trưng hữu ích cho tiếng Việt. Trình bày quá trình lựa chọn, xây dựng mô hình và tiến hành thực nghiệm một mô hình tích hợp tách từ và gán nhãn từ loại tiếng Việt

I KHÁI QUÁT VỀ TÁCH TỪ VÀ GÁN NHÃN TỪ LOẠI TIẾNG VIỆT

LI	Khái quát về tách từ tiếng Việt	. 1
1.2	Khái quát về gán nhãn từ loại - POS tagging	. 2
1.2.1	Giới thiệu về bài toán gán nhãn từ loại.....	. 2
1.2.2	Các hướng tiếp cận bài toán POS tagging	. 4
1.2.3	Các nghiên cứu gán nhãn từ loại cho tiếng Việt	. 6
1.3	Vấn đề tích hợp tách từ và gán nhãn từ loại	. 7
	MỘT MÔ HÌNH TÁCH TỪ TIẾNG VIỆT	9
2.1	Các mô hình liên quan	. 9
2.1.1	Mô hình dựa vào từ điển.....	. 9
2.1.2	Mô hình nhận dạng tên thực thể - Named Entity Recognition . .	. 10
2.1.3	Mô hình N-gram	. 10
2.2	Phân tích các mô hình	. li
2.3	Thiết kế tập đặc trưng	. 12
2.3.1	FS1: Đặc trưng trích từ mô hình tách từ dựa vào từ điển	. 13
2.3.2	FS2: Đặc trưng dựa vào mô hình nhận dạng tên thực thể	. 13
2.3.3	FS3: Đặc trưng dựa vào mô hình N-gram	. 14
2.4	Kết quả thực nghiệm	. 15
2.4.1	Đánh giá các đặc trưng FS1 và FS2 so với các mô hình trước đó .	. 15
2.4.2	Đánh giá tầm quan trọng của từng tập thuộc tính	. 16

2.5	Đánh giá kết quả tách từ.....	17
3	MỘT MÔ HÌNH GÁN NHÃN TỪ LOẠI TIẾNG VIỆT	19
3.1	Xây dựng corpus gán nhãn từ loại cho tiếng Việt	19
3.1.1	Thiết kế tập thể VnPOSTag	19
3.1.2	Mô tả bộ dữ liệu làm vnPOS corpus	21
3.1.3	Xây dựng vnPOS corpus.....	22
3.2	Gán nhãn từ loại bằng phương pháp Maximum Entropy Model	24
3.2.1	Mô hình xác suất	24
3.2.2	Các đặc trưng của gán nhãn từ loại	25
3.3	Đề xuất mô hình gán nhãn từ loại cho tiếng Việt	26
3.3.1	Gán nhãn từ loại dựa vào thông tin từ	27
3.3.2	Gán nhãn từ loại dựa vào hình vị	29
3.4	Đánh giá kết quả gán nhãn từ loại	30
4	MỘT MÔ HÌNH TÍCH HỢP TÁCH TỪ VÀ GÁN NHÃN TỪ LOẠI TIẾNG VIỆT	32
4.1	Các mô hình tích hợp cho tiếng Trung	32
4.2	Lựa chọn mô hình tích hợp cho tiếng Việt	34
4.3	Xây dựng mô hình và tiến hành thực nghiệm	35
4.3.1	Features.....	35
4.3.2	Giải mã	35
4.3.3	Kết quả.....	36
4.4	Thảo luận.....	36
A	Chú giải tập từ loại vnPOS	44