NGHIÊN CỨU VỀ GÁN NHÃN VAI NGHĨA CHO TIẾNG VIỆT

Nguyễn Quang Huy – NCS KHMT K21 – quanghuyinfor@gmail.com

I. TỔNG QUAN VỀ GÁN NHÃN VAI NGHĨA

Abstract - Gán nhãn vai nghĩa (Semantic Role Labeling - SRL) là bài toán xác định vai nghĩa cho các thành phần trong câu, làm rõ cấu trúc, mối quan hệ giữa vị từ và các tham đối cấu thành câu. Về bản chất, gán nhãn vai nghĩa chỉ là một bài toán trung chuyển, chứ không thật sự là nền tảng của một ứng dụng xử lý ngôn ngữ tự nhiên cụ thể nào. Tuy nhiên, đây lại là một bước cần thiết không thể bỏ qua trong các ứng dụng liên quan đến ngôn ngữ tự nhiên. Việc xác định đúng vai nghĩa của các thành phần trong câu là một vấn đề trung tâm của mọi hệ xử lý ngôn ngữ tự nhiên. Trong các nghiên cứu của mình, chúng tôi tiến hành định nghĩa một số vai nghĩa thông dụng, xây dựng kho ngữ liệu tiếng Việt có gán nhãn vai nghĩa. Bước đầu áp dụng phương pháp phân tích cú pháp đầy đủ để xây dựng công cụ gán nhãn vai nghĩa tự động cho tiếng Việt. Các thử nghiệm với tiếng Việt cho kết quả tương đối khả quan.

Keywords: Vai nghĩa; Gán nhãn vai nghĩa; Xử lý ngôn ngữ; Semantic Role Labeling.

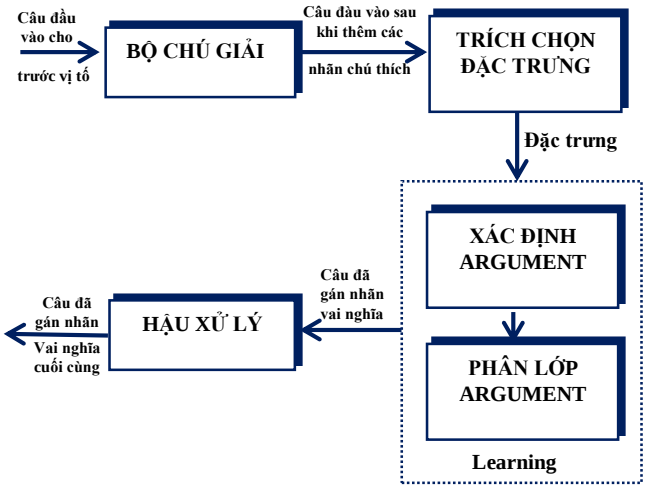
II. KẾT QUẢ NGHIÊN CỨU BÀI TOÁN SRL

Ngôn ngữ	P	R	F	Tác giả
Anh	81.18	74.92	77.92	Punyakanok-2005
Trung	81.03	72.38	76.46	Sun et al. - 2010
Ả-rập	75.23	71.65	72.20	D. M. Lundgren -2013
Nhật	86.74	88.48	87.61	S. Arora et al. – 2008

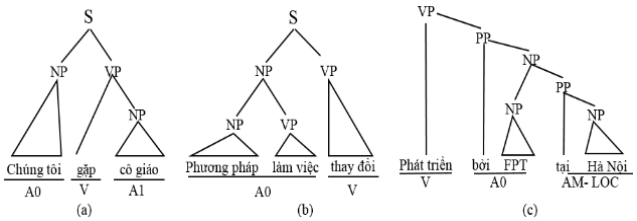
III. KHO NGỮ LIỆU CÓ GÁN NHÃN VAI NGHĨA

TT	Vai nghĩa	Nội dung
1	A0	Vai tác thể
2	A1	Vai bị thể
3	A2	Kẻ hưởng lợi, phương tiện, công cụ
4	A3	Điểm xuất phát
5	A4	Điểm kết thúc
6	AM-COM	Vai đối tượng đồng hành
7	AM-LOC	Vai vị trí của đối tượng, sự kiện,…
8	AM-TMP	Vai thời gian
9	AM-GOL	Vai mục đích, đích hướng tới
10	AM-MNR	Cách thức thực hiện hành động
11	AM-CAU	Vai nguyên nhân
12	AM- EXT	Mức độ tác động của hành động
13	AM- DIS	Vai kết nối
14	AM-NEG	Vai nghĩa phủ định
15	AM-DIR	Vai nghĩa chiều, hướng

IV. KIẾN TRÚC TỔNG QUAN HỆ THỐNG SRL



V. PHƯƠNG PHÁP PHÂN TÍCH CÚ PHÁP ĐẦY ĐỦ



Ánh xạ giữa nhãn vai nghĩa và các thành phần cú pháp: (a) ánh xạ 1-1 (A0 - NP); (b) ánh xạ 1-n cùng cha (A0 - NP,VP); (c) Ánh xạ 1-n khác cha (A0 - PP, NP)

VI. DỮ LIỆU THỬ NGHIỆM

Tệp	Training	Devel.	Test
Số câu	1.000	200	100
Động từ	1739	618	175
Tham đối	6066	1154	412

Dữ liệu	Traning 1.000 câu		Devel. 200 câu		Test 100 câu	
Vai nghĩa	**Số lượng**	**Tỷ lệ %**	**Số lượng**	**Tỷ lệ %**	**Số lượng**	**Tỷ lệ %**
A0	1586	26.15	282	24.44	139	33.74
A1	1984	32.71	376	32.58	115	27.91
A2	100	1.65	24	2.08	14	3.40
A3	18	0.30	5	0.43	0	0.00
A4	8	0.13	0	0.00	0	0.00
AM-COM	14	0.23	3	0.26	2	0.49
AM-LOC	376	6.20	94	8.15	10	2.43
AM-TMP	458	7.55	91	7.89	34	8.25
AM-GOL	74	1.22	11	0.95	14	3.40
AM-MNR	266	4.39	57	4.94	21	5.10
AM-CAU	42	0.69	8	0.69	7	1.70
AM- EXT	552	9.10	112	9.71	15	3.64
AM- DIS	266	4.39	33	2.86	23	5.58
AM-NEG	150	2.47	33	2.86	12	2.91
AM-DIR	172	2.84	25	2.17	6	1.46
Tổng số	**6066**		**1154**		**412**	

VII. KẾT QUẢ

Vai nghĩa	Precision	Recall	F1
Trung bình	**39,72**	**31,62**	**35,19**
A0	45,12	36,33	40,25
A1	42,27	30,78	35,62
A2	39,54	31,37	34,98
A3	0,00	0,00	0,00
A4	0,00	0,00	0,00
AM-COM	48,45	39,18	43,32
AM-LOC	43,25	33,41	37,70
AM-TMP	47,78	35,45	40,70
AM-GOL	37,12	27,54	31,62
AM-MNR	32,98	29,03	30,88
AM-CAU	34,11	27,55	30,48
AM- EXT	38,36	31,07	34,33
AM- DIS	38,27	30,74	34,09
AM-NEG	35,47	30,23	32,64
AM-DIR	33,63	28,41	30,80

VIII. KẾT LUẬN

Trong các nghiên cứu tiếp theo, chúng tôi sẽ tập trung xây dựng kho ngữ liệu tiếng Việt có chú thích vai nghĩa lớn hơn cả về số lượng và chất lượng. Sử dụng thêm một số đặc trưng riêng của tiếng Việt trong việc huấn luyện mô hình. Đồng thời tiến hành thử nghiệm một số phương pháp học máy mới nhằm nâng cao kết quả.