Báo cáo kỹ thuật
Xây dựng hệ thống tách từ tiếng Việt
underthesea v1.1.12

Vu Anh
underthesea
anhv.ict91@gmail.com

Tóm tắt

Trong báo cáo này, chúng tôi mô tả chương trình tách từ tiếng Việt, được tích hợp trong phiên bản underthesea phiên bản 1.1.12. Các công trình nghiên cứu trước đã rất thành công trong bài toán tách từ, chúng tôi muốn nghiên cứu lại sự hiệu quả của phương pháp Conditional Random Fields trong bài toán này. Sau đó xây dựng hệ thống tách từ hoàn chỉnh. Mã nguồn của chương trình được open source tại github.

1 Giới thiệu

Tách từ là một bài toán quan trọng trong việc xử lý rất nhiều ngôn ngữ. Đối với tiếng Việt, nhiệm vụ này khá khó khăn do một từ tiếng Việt thường gồm nhiều tiếng ghép lại. Ví dụ như từ *giáo viên* gồm hai tiếng *giáo* và *viên*.

Trong nghiên cứu này, chúng tôi xây dựng chương trình dựa trên giải thuật Conditional Random Fields trên bộ dữ liệu VLSP 2013.

2 Các công trình liên quan

Bài toán tách từ tiếng Việt đã được nghiên cứu từ khá lâu. Nguyen et al. (2006) sử dụng CRFs và SVM trên dữ liệu khoảng 7800 câu. Huyen et al. (2008) sử dụng các phương pháp kết hợp. Nguyen and Le (2016) sử dụng logistic regression trên 78000 câu.

3 Mô tả hệ thống

3.1 Hệ thống tách từ

Hệ thống tách từ trong underthesea được chia làm hai bước. Bước đầu tiên là bước tiền xử lý. Trong bước này, văn bản được tách câu và tokenize sử dụng regular expression. Bước thứ hai, các từ được biểu diễn dưới dạng một bài toán gán nhãn chuỗi.

3.2 Thuật toán Conditional Random Fields

Thuật toán Conditional Random Fields (CRFs) (Lafferty et al., 2001) được sử dụng đã tính toán xác suất của chuỗi đầu ra cho bởi chuỗi đầu vào. Xác suất của chuỗi trạng thái $S = \langle s_1, s_2, ..., s_T \rangle$ cho bởi quan sát $O = \langle o_1, o_2, ..., o_T \rangle$ được tính bởi công thức:

$$P(S|O) = \frac{1}{Z_o} \exp(\sum_{t=1}^{T} \sum_{k} \lambda_k x f_k(s_{t-1}, s_t, o, t))$$

trong đó, $f_k(s_{t-1}, s_t, o, t)$ làm một hàm đặc trưng ứng với trọng số λ_k, được học thông qua quá trình huấn luyện.

3.3 Features

Các đặc trưng được đề xuất

features	description
T[-2], T[-1], T[0], T[1], T[2]	unigram
T[-2,-1], T[-1,0], T[0,1], T[1,2]	bigram
T[-2,0], T[-1,1], T[0,2]	trigram
T[-1].isdigit, T[0].isdigit, T[1].isdigit	digit

4 Thực nghiệm

4.1 Dữ liệu

Để so sánh độ chính xác của chương trình. Chúng tôi sử dụng bộ dữ liệu đã được sử dụng trong Nguyen et al. (2018) và Nguyen and Le (2016). Dữ liệu huấn luyện gồm 75 nghìn câu được lấy từ dữ liệu huấn luyện của bài toán tách từ trong VLSP 2013. Dữ liệu kiểm thử gồm 2120 câu lấy từ bộ dữ liệu gán nhãn từ loại trong VLSP 2013.

4.2 Chỉ số đánh giá

Chúng tôi sử dụng precision, recall và f1 làm các chỉ số đánh giá.

$$F_1 = \frac{2PR}{P+R}$$

trong đó P (Precision), và R (Recall) được định nghĩa như sau:

$$P = \frac{NE_{\text{true}}}{NE_{\text{sys}}}$$

$$R = \frac{NE_{\text{true}}}{NE_{\text{ref}}}$$

với

NE_{true}: The number of NEs in gold data

NE_{sys}: The number of NEs in recognizing system

NE_{ref}: The number of NEs which is correctly recognized by the system

4.3 Kết quả

We conduct our experiment on VLSP 2013 dataset, the result show we achieve 97.3%

system	features	result
s1	ngram	96.42%
s2	s1 + lower	96.45%
s3	s2 + isdigit	96.54%
s4	s3 + istitle	96.45%
s5	s4 + unigram is in dict	96.45%
s6	s5 + bigram is in dict	97.34%
sn	full	97.31%

5 Kết luận

Trong báo cáo này, chúng tôi đã mô tả hệ thống tách từ được tích hợp trong underthesea phiên bản 1.1.12.

References

Nguyen Thi Minh Huyen, Azim Roussanaly, Hô Tuong Vinh, et al. 2008. A hybrid approach to word segmentation of vietnamese texts. In *Language and Automata Theory and Applications*, Springer, pages 240–249.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '01, pages 282–289. http://dl.acm.org/citation.cfm?id=645530.655813.

Cam-Tu Nguyen, Trung-Kien Nguyen, Xuan Hieu Phan, Le-Minh Nguyen, and Quang-Thuy Ha. 2006. Vietnamese word segmentation with crfs and svms: An investigation. In *PACLIC*.

Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras, and Mark Johnson. 2018. A fast and accurate vietnamese word segmenter. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*.

T. Nguyen and A. Le. 2016. A hybrid approach to vietnamese word segmentation. In *2016 IEEE RIVF International Conference on Computing Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*. pages 114–119. https://doi.org/10.1109/RIVF.2016.7800279.