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

Vu Anh
underthesea
anhv.ict91@gmail.com

Abstract

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. 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*.

2 Các công trình liên quan

Bài toán tách từ tiếng Việt đã được nghiên cứu từ khá lâu.

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 $S =< s_1, s_2, ..., s_T >$ cho bởi quan sát $O =< o_1, o_2, ..., o_T >$ 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

We propose conditional random fields for this problem.

Our final features

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 height

4 Thực nghiệm

4.1 Data sets

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 Evaluation Measures

We used Precision, Recall, F1 score as evaluation measures.

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

where P (Precision), and R (Recall) are determined as follows:

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

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

where

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

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

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

4.3 Kết quả

We conduct our experiment in VLSP 2013 dataset, the result show we archive 97.3%

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

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

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.