Báo cáo kỹ thuật
Xây dựng chương trình tách từ tiếng Việt
underthesea v1.1.12

Vu Anh
underthesea
anhv.ict91@gmail.com

Abstract

In this report, we describe our word segmentation system for Vietnamese, which is integrated in underthesea version 1.1.8. Our system is open-source and available at https://github.com/undertheseanlp/word_tokenize

1 Giới thiệu

Word Segmentation is an important task in many language. In Vietnamese, it is more difficult because one word can contains two and three syllables.

2 System Description

2.1 Conditional Random Fields

In order to solve word segmentation problem, there are many algorithms such as HMM, SVM, Riple Down Rules. In our experiments, we use condtional random fields, which yields many success for sequence labeling problem.

In this session, we brife describe conditional random fields algorithm.

2.2 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

3 Evaluation

3.1 Data sets

To be updated

3.2 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

3.3 Results

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

4 Conclusion

We have introduced our approach and its experimental result in word segmentation for Vietnamese text.

References