Vietnamese Text Classification System in underthesea v1.1.8

Vu Anh
underthesea
Hanoi, Vietnam
anhv.ict91@gmail.com

Abstract

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

1 Introduction

Text classification is an important task in Natural Language Processing with many applications, such as web search, information retrieval, ranking and document classification.

In recent years, natural language processing and text classification have had a lot of works with encouraging results of research comminity in side and outsite Vietnam. The relevant works outside Vietnam have been published a lot.

2 System Description

We takes some experiments in Logistic Regression, Naive Bayes, SVM (Suykens and Vandewalle, 1999), fasttext models (Joulin et al., 2016).

2.1 SVM

SVM algorithm works by creating support vectors that separating two classes in n-dimensional space, where each dimension is represented by a feature. The separation is done by finding the largest separation margin between the features from the two classes and a vector.

2.2 FastText

FastText maps each vocabulary to a real-valued vector, with unknown words having a special vocabulary ID. A document can be represented as the average of all these vectors. Then FastText will train a maximum entropy multi-class classifier on the vectors and the output labels.

3 Evaluation

3.1 Data sets

VNTC (Hoang et al., 2007): A Vietnamese corpus based on the four largest circulation Vietnamese online newspapers: VnExpress, TuoiTre Online, Thanh Nien Online, Nguoi Lao Dong Online. The collected texts are automatically preprocessed (removing the HTML tags, spelling normalization) by Teleport software and various heuristics. There followed a stage of manual correction by linguists (five master students in Linguistics of University of Social Sciences, VNU-HCM city, Vietnam) who reviewed and adjusted the documents which are classified to the wrong topics. Final relatively large and sufficient corpus includes about 100,000 documents.

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:

$$P = \frac{\text{the text number classified by the model correctly}}{\text{the text number classified correctly in practice}}$$

$$R = \frac{\text{the text number classified by the model correctly}}{\text{the text number classified by the model}}$$

3.3 Results

Table 3.3 show result on each system in VNTC dataset.

Method	P	R	F1
Fasttext	0.86	0.86	0.85

Table 1: Result

4 Conclusions

There are a lot of thing you we to try in next version. We want to explore text classification tutorial of Google[1], which has more details to build text classification system.

References

Vu Cong Duy Hoang, Dien Dinh, Nguyen Le Nguyen, and Hung Quoc Ngo. 2007. A comparative study on vietnamese text classification methods. In *2007 IEEE International Conference on Research, Innovation and Vision for the Future in Computing & Communication Technologies, RIVF 2007, Hanoi, Vietnam, 5-9 March 2007*. pages 267–273. https://doi.org/10.1109/RIVF.2007.369167.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *CoRR* abs/1607.01759. http://arxiv.org/abs/1607.01759.

J.A.K. Suykens and J. Vandewalle. 1999. Least squares support vector machine classifiers. *Neural Processing Letters* 9(3):293–300. https://doi.org/10.1023/A:1018628609742.

[1] Google Text Classification Tutorial