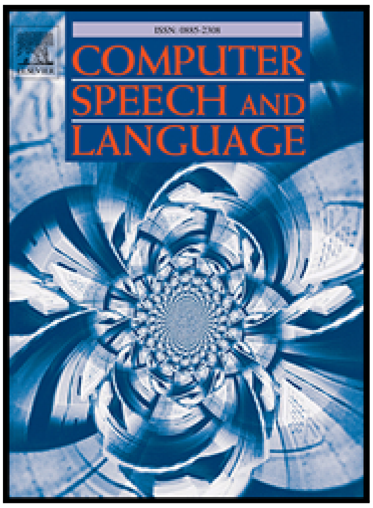

An Empirical Study on POS Tagging for Vietnamese Social Media Text

Ngo Xuan Bach, Nguyen Dieu Linh, Tu Minh Phuong

PII: S0885-2308(17)30027-X
DOI: 10.1016/j.csl.2017.12.004
Reference: YCSLA 904

To appear in: *Computer Speech & Language*

Received date: 6 February 2017
Revised date: 14 August 2017
Accepted date: 13 December 2017

Please cite this article as: Ngo Xuan Bach, Nguyen Dieu Linh, Tu Minh Phuong, An Empirical Study on POS Tagging for Vietnamese Social Media Text, *Computer Speech & Language* (2017), doi: 10.1016/j.csl.2017.12.004

An Empirical Study on POS Tagging for Vietnamese Social Media Text

Ngo Xuan Bach[a,b,*], Nguyen Dieu Linh[a], Tu Minh Phuong[a,b]

[a]*Department of Computer Science, Posts and Telecommunications Institute of Technology, Vietnam*
[b]*Machine Learning & Applications Lab, Posts and Telecommunications Institute of Technology, Vietnam*

Abstract

Part-of-speech (POS) tagging is a fundamental task in natural language processing (NLP). A robust POS tagger plays an important role in most NLP problems and applications, including syntactic parsing, semantic parsing, machine translation, and question answering. Although a lot of efficient POS taggers has been developed for general, conventional text, little work has been done for social media text. In this paper, we present an empirical study on POS tagging for Vietnamese social media text, which shows several challenges compared with tagging for general text. Social media text does not always conform to formal grammars and correct spelling. It also uses abbreviations, foreign words, and emoticons frequently. A POS tagger developed for conventional text would perform poorly on such noisy data. We address this problem by proposing a tagging model based on Conditional Random Fields (CRFs) with various kinds of features for Vietnamese social media text. We also investigate the effect of features extracted from word clusters under the Brown and canonical correlation analysis (CCA) based clustering in semi-supervised settings. We introduce an annotated corpus for POS tagging, which consists of more than four thousand sentences from Facebook, the most popular social network in Vietnam. Using this corpus, we performed a series of experiments to evaluate the proposed model. Our model achieved 88.26% and 88.92% tagging accuracy in supervised and semi-supervised scenarios respectively, which are nearly 12% improvement over vnTagger, a state-of-the-art and most widely used Vietnamese POS tagger developed for general, conventional text. In addition, the semi-supervised model outperformed, in terms of accuracy, the version of vnTagger trained on the same Facebook dataset, showing the usefulness of word cluster features[1].

Keywords: Part-of-Speech tagging, Social media text, Conditional Random Fields, Word Clustering

[1] This paper is an improved and extended version of Bach et al. (2016)

[*] Corresponding author; tel: +84 898980268

Email addresses: bachnx@ptit.edu.vn (Ngo Xuan Bach), ndlinh@ptit.edu.vn (Nguyen Dieu Linh), phuongtm@ptit.edu.vn (Tu Minh Phuong)