Fundamental Tools and
Available for Vietnames

Kanji Takahashi, Kazuhide Yamamoto
Nagaoka University of Technology

Abstract

This paper presents our work on developing Vietnamese fundamenta
word segmentation and part-of-speech tagging, diacritics restoration
been either not publicly available so far or not attaining sufficient perfo
the tools to the public, in both software packages and web tools. For
achieved high accuracy. We briefly present the tasks, the methods ar

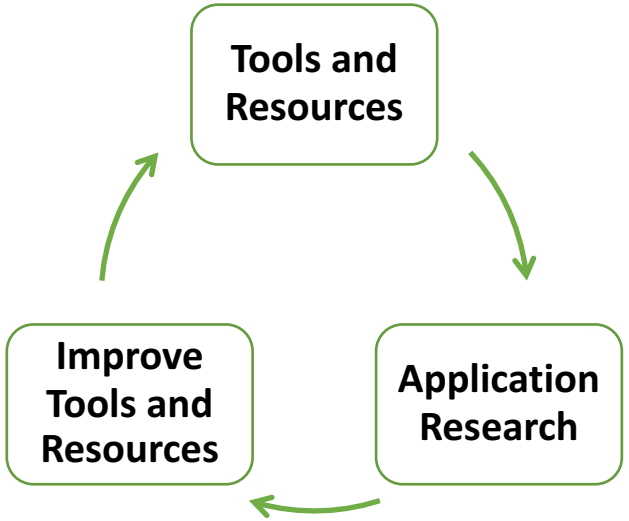
Resource are
e Analysis

al tools and a resource for analysis. These tools are for
n, and orthographical variants dictionary. All of them have
formance. We have developed the tools and released
r development, we utilize state-of-the-art methods and
nd the performance of each tool and resource.

Introduction

Vietnamese is a low resource language.

Basic tools and resources are important to improve NLP research.



Fundamental tools and resources contribute to development of NLP research like eco system.

Vietnamese Language

Isolated Language

Like Chinese.
Tôi là sinh viên. (I am a student.)

Vietnamese Alphabets

Like Chinese pinyin
Sequence of syllables

sinh nhật
生　日
(birthday)

Over 70 Million Speakers

Vietnamese NLP technology helps a lot of people.

Joint Word Segmentation and POS Tagging Tool

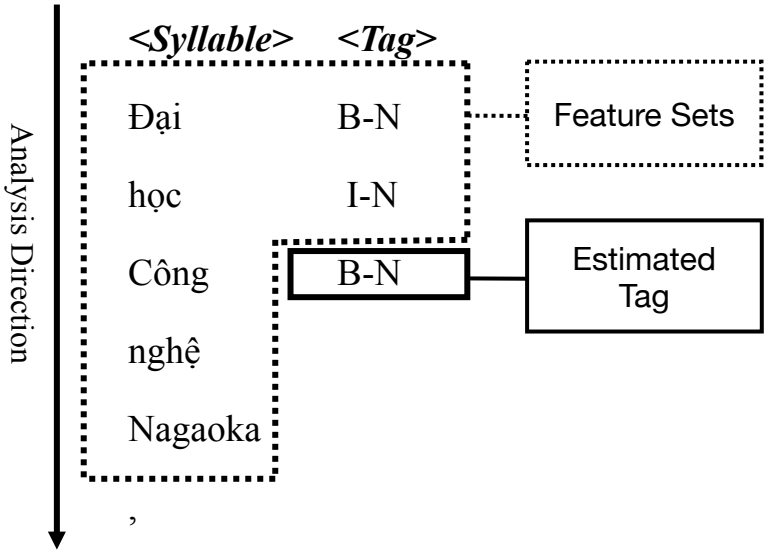
Method

Joint word segmentation and POS tagging using SVM or CRF.

IOB2 tag with POS are tagged.

vnPOS corpus,10-cross-validation

- 6,962 sentences
- 183,398 syllables
- 144,010 words
- 15 POS tags

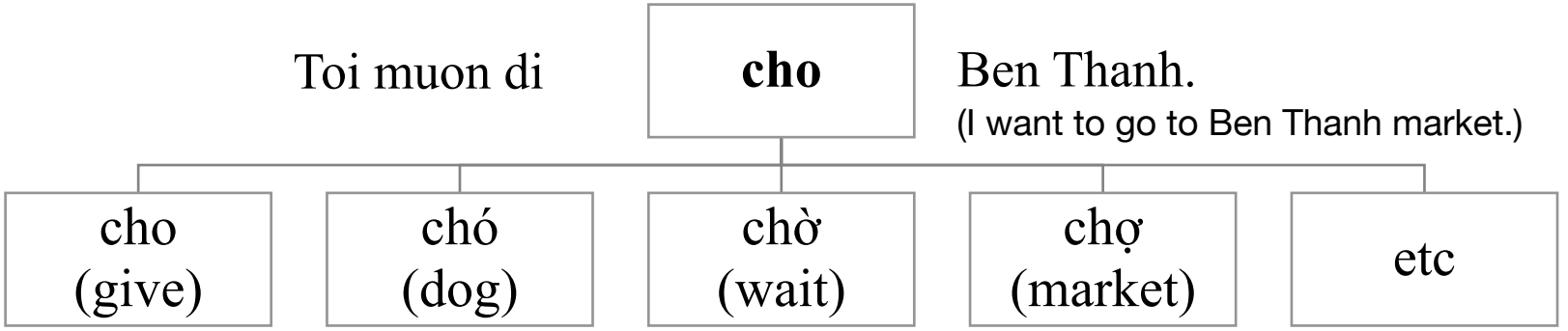


Method	Recall(%)	Precision(%)	F-Value(%)
SVM(Yamcha)	87.83	89.24	88.53
CRF(CRF++)	89.10	90.10	89.60

SVM https://github.com/kanjirz50/viet-morphological-analysis-svm
CRF https://github.com/kanjirz50/viet-morphological-analysis-crf

Diacritics Restoration Tool

In some Vietnamese text, diacritics are dropped.



Tôi muốn đi chợ Bến Thành.

We adopt a state-of-the-art method based on a point-wise prediction approach*.

83% accuracy.

https://github.com/kanjirz50/restore-tonemark

* Tuan Anh Luu, and Kazuhide Yamamoto, A pointwise approach for Vietnamese diacritics restoration, *IALP2012*, pp.189– 192, 2012

Syllable Normalization Dictionary

Diacritics mark causes orthographical variation.

Syllable 1	Frequency 1	Syllable 2	Frequency 2
thủy	14,280	thuỷ	2,307
thúy	6,047	thuý	912
khóe	1031	khoé	409

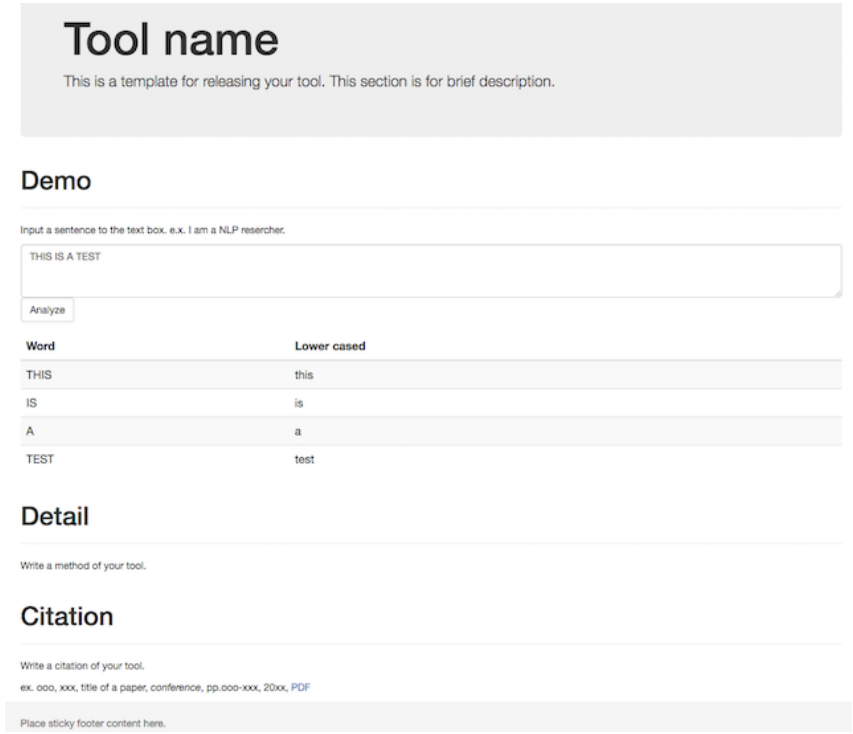
These syllables are Collected from 1.7 million sentences.

Unicode Problem

- ”ắ” can be “U+1eb5” or “U+0103” + “U+0303”

We constructed normalization dictionary and preprocessing scripts.

Web demonstration template for NLP research



NLP researchers should open own work on the Internet as a demo.

This template helps you release your research.

Of course, our tools uses this template.

https://github.com/kanjirz50/web-nlp-interface

For advance in NLP research, it is important to provide some fundamental analyzers as public and share.