A Hybrid Approach to Vietnamese Word Segmentation

Student: Nguyen Tuan Phong
Instructors: Assoc. Prof. Le Anh Cuong
Dr. Le Nguyen Khoi
Faculty of Information Technology, VNU University of Engineering and Technology

Introduction

- Word Segmentation (WS) is the very first task for Natural Language Processing (NLP) in Vietnamese.
- WS is the task to detect boundary of words in an input text.
- Input: raw text
- Output: word-segmented text
- Example:
  Học sinh học sinh học.
  → *Học_sinh học sinh_học .*

("_" is separator of syllables inside a word)

- Common approaches:
  - Dictionary-based approach
  - Statistical approach
  - Hybrid approach → state-of-the-art results

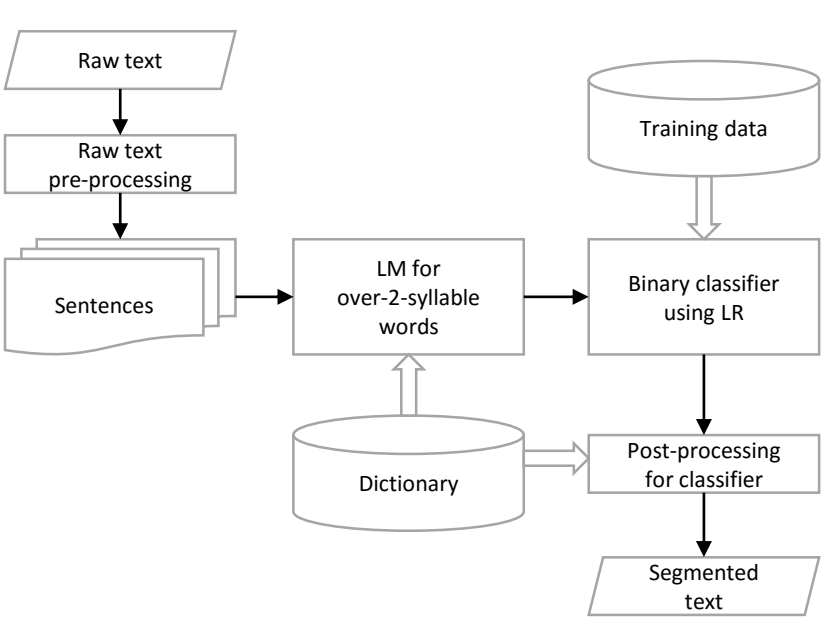
Difficulties

- Vietnamese words are made of one or more syllables.
- Most of words are 2-syllable ones.
- **Overlap ambiguity**
  - syllable sequence: $s_i s_{i+1} s_{i+2}$
  - both $s_i s_{i+1}$ and $s_{i+1} s_{i+2}$ are words in dictionary
  - only one word is right in current context
- **Proper names:** two proper names appear consecutively. Example:
  *Tổng_thống **Mỹ Barack_Obama** đang thăm Việt_Nam .*
- **Out-of-vocabulary:** words have not appeared in dictionary.

Proposed approach

Motivation

- Low frequency of over-2-syllable words → just use longest matching (LM) algorithm to cover them.
- Binary classifier for white spaces
- Using logistic regression (LR) for binary classifier → simple to have post-processing



Architecture of our word segmentation system

Feature set for logistic regression

No.	Template
1	$(f_i), i = -2, -1, 0, 1, 2$
2	$(f_i, f_{i+1}), i = -2, -1, 0, 1$
3	$(t_i), i = -2, -1, 0, 1, 2$
4	$(t_i, t_{i+1}), i = -2, -1, 0, 1$ && $t_i \neq LOWER$
5	$(t_i, t_{i+1}, t_{i+2}), i = -2, -1, 0$ && $t_i \neq LOWER$
6	$(t_0 = t_1 = LOWER$ && $f_0 = f_1)$?
7	$(t_0 = t_1 = UPPER$ && $isVNFamilyName(s_0))$?
8	$(t_0 = t_1 = UPPER$ && $isVNSyllable(s_0)$ && $!isVNSyllable(s_1))$?

- f is lowercase-simplified form of syllable s
- t is syllable type (LOWER, UPPER, ALLUPPER, NUMBER, OTHER)

Post-processing

- Use dictionary to verify the predictions that have low confidence conducted by the classifier
- Use dictionary to verify 3-syllable words

Experiments

Accuracy of sub-systems (10-fold CV on Vietnamese Treebank of 75k sentences)

Sub-system	P	R	F
LM	97.11	97.31	97.21
LR	97.95	98.29	98.12
LM + LR	98.11	98.16	98.14
LR + Post	98.59	98.99	98.79
LM + LR + Post	98.77	98.87	98.82

Comparison to other tools (10-fold CV on Vietnamese Treebank of 75k sentences)

Toolkit	P	R	F
vnTokenizer	97.61	96.86	97.23
JVnSeg-MaxEnt	97.18	97.28	97.23
JVnSeg-CRFs	97.58	97.68	97.63
DongDu	97.44	98.01	97.72
Ours	**98.77**	**98.87**	**98.82**

Segmentation speed (measurement on a corpus of 1k articles)

Toolkit	Speed (tokens/s)
JVnSeg-CRFs	764
JVnSeg-MaxEnt	1082
vnTokenizer	5322
DongDu	16709
Ours	**33705**

- Our system provides the most accurate results for Vietnamese word segmentation while evaluating on Vietnamese Treebank corpora of 75k word-segmented sentences.
- Our system runs faster than any other current toolkit.

Publication

- Toolkit: UETsegmenter (https://github.com/phongnt570/UETsegmenter)
- A paper submitted to the ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP):
  Phong Tuan Nguyen and Cuong Anh Le. 2016. *A Hybrid Approach to Vietnamese Word Segmentation*.