On the Effect of Part-of-Speech Information for Vietnamese Word Segmentation

Van-Hai Nguyen　Kanji Takahashi　Kazuhide Yamamoto
Nagaoka University of Technology
{hai, takahashi, yamamoto}@jnlp.org

1 Introduction

We present in this paper an investigation on the effect of part-of-speech (POS) information for Vietnamese word segmentation. According to the public resources of Building Basic Resources and Tools for Vietnamese Language and Speech Processing (VLSP), Vietnamese is classified as low resource language. Therefore, we focus on expanding resources. Among these, word segmentation is the most basic and important task, although the accuracy is not high as expected for general domains.

Vietnamese belongs to Asian language and isolated language such as Chinese and Thai. Vietnamese writing system uses Latin character with additional diacritics for tones. Thus, to correctly determine the word, we must resolve word boundaries and choose the correct word for this sentence. POS information is a category of the word which have similar grammatical properties, which is helpful for resolving word ambiguity and improving word segmentation.

Almost existing Vietnamese analyzer adopt two-step analysis. First step is word segmentation. Next is POS tagging for word segmented text. We propose using POS information for improving Vietnamese word segmentation. We use a CRF based tools[1] for investigation . This tool bases CRF, joint word segmentation and POS tagging as a chunking task.

2 Vietnamese Word

2.1 Vietnamese Morpheme

In Vietnamese morpheme is the basic unit, separated by space. It can be identified by the human who understands Vietnamese and by computer automatically. One morpheme can be:

- Single word: "tôi" (I), "chị" (sister)
- Component of compound word: "hoa" (flower), "hồng" (pink) and a word "hoa hồng" (rose)

2.2 Vietnamese Word

Vietnamese word consists of one or more than morphemes. Based on the structure of words, we classify Vietnamese word into three categories: single word, compound word and duplicative words.

- Single word contains one morpheme that implies meaning. For example: "tôi" (I), "chị" (sister)
- Compound word is words that consist of more than one morpheme. For example: "suy nghĩ" (think), "mặt trời" (sun), "ẩm ướt" (wet)
- A word may have duplication of its morpheme or that with a little phonic difference. It is called as reduplication. For example: "người người" (everyone), "phinh phính" (plump)

3 Vietnamese Word Segmentation

3.1 Difficulty and Challenge

The Vietnamese writing system is based on Latin characters, which present the pronunciation but not the meaning of the words, Vietnamese word segmentation faces on the ambiguity problem of the word.

Especially, the same surface form but different word segmentation appear in Vietnamese.

For example “thông tin” has two different word segmentation: thông (pine) tin (news) and thông tin (information). In other case “quần áo” has two different word segmentation: quần (trouser) áo (shirt) and quần áo (clothes).

So, in the sentences “tốc độ truyền thông tin ngày càng cao” both of word “truyền thông” and “tin thông” are words. Solving this kind of ambiguity can improve the Vietnamese word segmentation.

3.2 Existing Work

vnTokenizer is developed by Le [2], which implements a hybrid approach to automatically tokenize Vietnamese text. The approach combines both finite-state automata technique, regular expression parsing and the maximum matching strategy which is augmented by statistical methods to resolve ambiguities of segmentation. The system has good F-score, about 95%.

Luu and Yamamoto develop DongDu using Pointwise method based on SVM. They treat word segmentation as a classification problem. The system has high accuracy around 98%.

Takahashi and Yamamoto developed SVM and CRF based tools using joint word segmentation and POS tagging as a chunking task. SVM and CRF tools are developed by Kudo.

4 Experiment

As mention in Section 3.1 the complexity of word structure and word boundary in Vietnamese causes ambiguity in word segmentation.

Our goal in this study is to improve word segmentation using POS information. For this purpose, we investigate the influence of POS information on word segmentation. It helps to know the difference between a word segmentator with and without POS information. In this experiment, we follow Takahashi's experiment, using CRF++[1]. We use POS information which is built by Oanh with 15 tags [2].

4.1 Dataset

The corpus contains 6962 sentences. We divided the corpus 6264 sentences for training and 696 sentences for testing.

For word segmentation with POS information, IOB2 tag is suitable to solve the problem as tagging. We use B tag and I tag to represent word boundary.
Tấp B-JJ
nập I-JJ
sắm B-VB
...
In word segmentation without POS information, IOB2 tags do not involve POS information.
Tấp B
nập I
sắm B
...

4.2 Results and Discussion

In this section, we discuss the results and the effect of POS information. Table 1 presents as sample of the results for the Vietnamese word segmentation with CRF++ classifier.

Table 1: **Sample of results**

Sent	Train	POS	noPOS
Hay	B	B-VB	B
đến	B	B-IN	B
nỗi	I	B-NC	B
một	B	B-D	B
nhóm	B	B-NC	B
thanh	B	B-NN	B
niên	I	I-NN	I
người	B	B-NC	B
Mỹ	B	B-NP	B

4.2.1 POS information improves word segmentation

We check the results manually and found out that using POS information is better than not using POS information in the experiments presented in Table 2. The *correct* column shows the correct answer of a word (B or B, I or B, I, I) in the sentence. The

[1] https://taku910.github.io/crfpp/

POS and *noPOS* column present the results of the prediction using POS information and not using POS information.

Table 2: **Sample of correct words segmented by using POS information**

Word	Correct	POS	noPOS
hết	B	B-AD	B
sức	I	I-AD	B
ngân	B	B-NN	B
hàng	I	I-NN	I
dữ	I	I-NN	B
liệu	I	I-NN	I
rượu	B	B-NN	B
bia	I	I-NN	B
dân	B	B-NN	B
sự	I	I-NN	B
Hoàng	B	B-NP	B
Mộng	I	I-NP	B
Hà	I	I-NP	I

The result of Table 2 shows some words that are predicted incorrectly by the system without using POS information. However, the system predicted the output correctly when using POS information. For example, the person name "bà Hoàng_Mộng_Hà" is correct answer, but without using POS information, the word is separated into "bà Hoàng Mộng_Hà".

Similarly, the correct answer "ngân_hàng_dữ_liệu" is separated into "ngân_hàng dữ_liệu" when the system does not use POS information.

According to Takahashi and Yamamoto, the previous tags in sentences, syllables and POS information, will be used as features to identify the next tag.

4.2.2 POS information does not improve word segmentation

However, we also have some results that are wrong in both approaches. Conjunction almost appear in this case. For example: "mặc cho", "đến nỗi", "hết sức", "mặc cho" and a difficult pronoun such as this word "Trường Đại_học Mở_Bán_công", both of approaches output the wrong segmentation.

And, some are correct without using POS information, but they are predicted wrong by the system with POS information which we show in Table 2. For example: "cả nĩa", "Giá dầu". Different with the isolated language such as Japanese, Vietnamese has special morpheme use to present time, space, or number..., and when standing alone, they do not have meaning. For example: "đã" presents for the action in the past. "tôi đã ăn cơm." (I ate meat.). They do not have any them features to classy in the same group. So, we need to know more details about the system how to use their features. In this method, sentence capture from the head to the end for only one direction. So, the next word will be predicted based on the estimated tags, syllables and POS information.

Table 3: **Sample of correct words are segmented which do not use POS information**

Word	Correct	POS	noPOS
đến	B	B-IN	B
nỗi	I	B-NC	B
cả	B	B-PP	B
nĩa	B	I-PP	B
Mặc	B	B-VB	B
cho	I	B-IN	B
giá	B	B-NN	B
dầu	B	I-NN	B

5 Conclusion and Future work

The results show POS information influence on Vietnamese Word Segmentation. We experimented with and without using POS infomation. The prediction accuracy is improved when using POS information as a feature. This opens a new approach to use POS information to improve the existing Vietnamese Word Segmentation.

Japanese morphological analyzers are successful by using lattice information based on the whole sentence. In the future, we want to adapt similar method for Vietnamese word segmentation.

Reference

[1] Kanji Takahashi and Kazuhide Yamamoto *Fundamental Tools and Resource are Available for*

Vietnamese Analysis. 2016 International Conference on Asian Language Processing (IALP), pp.246-249, 2016.

[2] Hong Phuong Le, Thi Minh Huyen Nguyen, Azim Roussanaly, Tuong Vinh Ho *A Hybrid Approach to Word Segmentation of Vietnamese Texts*. 2nd International Conference on Language and Automata Theory and Applications, pp.240-249, 2008.

[3] Tuấn Anh Lưu, and Kazuhide Yamamoto *Ứng dụng phương pháp Pointwise vào bài toán tách từ cho tiếng Việt*. http://viet.jnlp.org/dongdu

[4] C.T. Nguyen, T.K Nguyen, X.H. Phan, L.M. Nguyen, Q.T Ha. *Vietnamese Word Segmentation with CRFs and SVMs: An Investigation*. Proceedings of the 20th PACLIC, Wuhan, China, p.215-222, 2006.

[5] Tatsumi Yoshida, Kiyonori Ohtake, and Kazuhide Yamamoto, *Performance Evaluation of Chinese Analyzers with Support Vector Machines*. Journal of Natural Language Processing, vol.10(1), pp.109–131, 2003.