

# Underthesea 1.3.5 - Text Normalization

Vũ Anh <anhv.ict91@gmail.com>

## 1 Introduction

Text normalization is a fundamental task in Vietnamese natural language processing. There are some research in this topic (Trang et al., 2019). In this study, we focus on building a simple module for this task. We propose the method of using a set of rules to solve this problem. Then compare the effectiveness with some popular tools in Vietnamese. The results of the study are quite positive. The code is available at <https://github.com/undertheseanlp/underthesea>.

## 2 Vietnamese Text Normalization Task

### 2.1 Vietnamese Alphabet

According to wikipedia<sup>1</sup>, Vietnamese alphabet contains 29 letters and five diacritics used to designate tone.

There are 12 vowels and 17 consonants. Single Vowels (nguyên âm đơn) are A, Ă, Â, E, Ê, I, O, Ô, Ơ, U, Ừ, Y. Double vowels (nguyên âm đôi) are AI, AO, AU, ÂU, AY, ÂY, EO, ÊU, IA, IÊ, IU, OA, OĂ, OE, OI, ÔI, ƠI, OO, ÔÔ, UA, UĂ, UÂ, UÁ, UÊ, UI, Ừ, UO, UÔ, UƠ, UỞ, UU, UY, YÊ. Triple Vowels (nguyên âm ba) are IÊU, OAI, OAO, OAY, OEO, UAO, UÂY, UÔI, ƯƠI, ƯỚU, UYA, UYÊ, UYU, YÊU. Single consonants (phụ âm đơn) are B, C, D, Đ, G, H, K, L, M, N, P, Q, R, S, T, V, X. Compound consonants (phụ âm ghép) are CH, GI, GH, KH, NH, NG, NGH, PH, TH, TR, QU.

### 2.2 Text Normalization Tasks

Normalization tasks:

- Character normalization
- Punctuation standardization (luý → lúy, cứ → cừ)

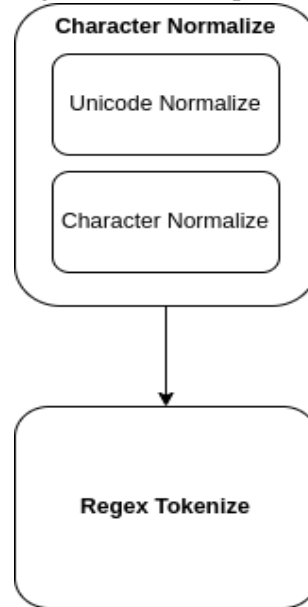
<sup>1</sup>[https://en.wikipedia.org/wiki/Vietnamese\\_alphabet](https://en.wikipedia.org/wiki/Vietnamese_alphabet)

Rules for placing accent marks in the Vietnamese text is described in the wikipedia article <sup>2</sup>

## 3 Methods

We construct 640 rules for mapping syllables, each rule is mapping between the incorrect syllable with the correct one.

Figure 3 show our process to normalize text.



## 4 Results

In order to evaluate the quality of our text normalization module, we compare our module with three popular tools benchmark in Vietnamese viet\_text\_tools (vtt) <sup>3</sup>, Vietnamese Text Normalizer (vtm) <sup>4</sup>, NLPUtils <sup>5</sup>

<sup>2</sup>[https://vi.wikipedia.org/wiki/Quy\\_t%E1%BA%AFC\\_%C4%91%E1%BA%B7t\\_d%E1%BA%A5u\\_thanh\\_trong\\_ch%E1%BB%AF\\_qu%E1%BB%91c\\_ng%E1%BB%AF](https://vi.wikipedia.org/wiki/Quy_t%E1%BA%AFC_%C4%91%E1%BA%B7t_d%E1%BA%A5u_thanh_trong_ch%E1%BB%AF_qu%E1%BB%91c_ng%E1%BB%AF)

<sup>3</sup>[https://github.com/enricobarzetti/viet\\_text\\_tools](https://github.com/enricobarzetti/viet_text_tools)

<sup>4</sup><https://github.com/langmaninternet/VietnameseTextNormalizer>

<sup>5</sup><https://gist.github.com/nguyenvanhieuvn/72ccf3ddf7d1>

	vtt	vtm	NLPUtils
Differences	1160	235	1408
Non miss spell	585	107	973
Miss spell	575	128	435

Table 1: The differences between our module with others

	Accuracy
Underthesea	1.0
Vietnamese Text Normalization	1.0
Viet Text Tools	1.0
NLPUtils	0.9992

Table 2: Evaluation in UD\_Vietnamese-UUD dataset

Table 1 shows the differences between our module with others

We release a dataset UD\_Vietnamese-UUD version 1.0.1-alpha contains 1145 sentences in Universal Dependency format.

Table 2 show the accuracy of four tools in UD\_Vietnamese-UUD version 1.0.1-alpha dataset

## 5 Discussion & Conclusions

In this paper, we conduct a fundamental research about Vietnamese Text Normalization. We build a `text_normalize` module for Vietnamese and release a dataset for evaluation.

In the future, we will develop a bigger dataset for better evaluation.

## References

Nguyen Thi Thu Trang, Dang Xuan Bach, and Nguyen Xuan Tung. 2019. A hybrid method for vietnamese text normalization. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, pages 104–109.

## 6 Appendices

In this section, we shows detail tables about five diacritics in Vietnamese

We follows to Vietnamese Collation Chart described in vietunicode<sup>6</sup>

<sup>6</sup><http://vietunicode.sourceforge.net/charset/v3.htm>

Diacritic	Examples
Ngang	ta
Huyền	tà
Hỏi	tả
Ngã	tã
Sắc	tá
Nặng	tạ

Table 3: Five diacritics

a	A	à	À	ă	Ă	ã	Ã	á	Á	ạ	Ạ
0061	0041	00E0	00C0	1EA3	1EA2	00E3	00C3	00E1	00C1	1EA1	1EA0
ă	Ă	ằ	Ằ	Ẳ	Ẵ	ẵ	Ẳ	ẵ	Ẳ	ặ	Ặ
0103	0102	1EB1	1EB0	1EB3	1EB2	1EB5	1EB4	1EAF	1EAE	1EB7	1EB6
â	Â	ầ	Ầ	ẫ	Ẫ	ẫ	Ẫ	ấ	Ấ	ậ	Ậ
00E2	00C2	1EA7	1EA6	1EA9	1EA8	1EAB	1EAA	1EA5	1EA4	1EAD	1EAC
b	B										
0062	0042										
c	C										
0063	0043										
d	D										
0064	0044										
đ	Đ										
0111	0110										
e	E	è	È	ẻ	Ẻ	ẽ	Ễ	é	É	ẹ	Ẹ
0065	0045	00E8	00C8	1EBB	1EBA	1EBD	1EBC	00E9	00C9	1EB9	1EB8
ê	Ê	ề	Ề	ể	Ễ	ễ	Ễ	ê	Ê	ệ	Ệ
00EA	00CA	1EC1	1EC0	1EC3	1EC2	1EC5	1EC4	1EBF	1EBE	1EC7	1EC6
f	F										
0066	0046										
g	G										
0067	0047										
h	H										
0068	0048										
i	I	ì	Ì	ỉ	Ỉ	ĩ	Ĩ	í	Í	ị	Ị
0069	0049	00EC	00CC	1EC9	1EC8	0129	0128	00ED	00CD	1ECB	1ECA
j	J										
006A	004A										
k	K										
006B	004B										
l	L										
006C	004C										
m	M										
006D	004D										
n	N										
006E	004E										
o	O	ò	Ò	ỏ	Ỏ	õ	Õ	ó	Ó	ọ	Ọ
006F	004F	00F2	00D2	1ECF	1ECE	00F5	00D5	00F3	00D3	1ECD	1ECC
ô	Ô	ồ	Ồ	ố	Ổ	ỗ	Ỗ	ố	Ổ	ộ	Ộ
00F4	00D4	1ED3	1ED2	1ED5	1ED4	1ED7	1ED6	1ED1	1ED0	1ED9	1ED8
ơ	Ơ	ờ	Ờ	ở	Ở	ỡ	Ỡ	ớ	Ớ	ợ	Ợ
01A1	01A0	1EDD	1EDC	1EDF	1EDE	1EE1	1EE0	1EDB	1EDA	1EE3	1EE2

Table 4: Vietnamese Collation Chart 1

<b>p</b> 0070	<b>P</b> 0050										
<b>q</b> 0071	<b>Q</b> 0051										
<b>r</b> 0072	<b>R</b> 0052										
<b>s</b> 0073	<b>S</b> 0053										
<b>t</b> 0074	<b>T</b> 0054										
<b>u</b> 0075	<b>U</b> 0055	ù 00F9	Û 00D9	ủ 1EE7	Ũ 1EE6	ũ 0169	Ũ 0168	ú 00FA	Ú 00DA	ụ 1EE5	Ụ 1EE4
<b>ư</b> 01B0	<b>Ư</b> 01AF	ừ 1EEB	Û 1EEA	ử 1EED	Ũ 1EEC	ũ 1EEF	Ũ 1EEE	ứ 1EE9	Ứ 1EE8	ự 1EF1	Ự 1EF0
<b>v</b> 0076	<b>V</b> 0056										
<b>w</b> 0077	<b>W</b> 0057										
<b>x</b> 0078	<b>X</b> 0058										
<b>y</b> 0079	<b>Y</b> 0059	ỳ 1EF3	Ỡ 1EF2	ỷ 1EF7	Ỡ 1EF6	ỹ 1EF9	Ỡ 1EF8	ý 00FD	Ỡ 00DD	ỵ 1EF5	Ỡ 1EF4
<b>z</b> 007A	<b>Z</b> 005A										

Table 5: Vietnamese Collation Chart 2