Abstract
This paper presents our work on developing Vietnamese fundamental tools and a resource for analysis. These tools are for word segmentation and part-of-speech tagging, diacritics restoration, and orthographical variants dictionary. All of them have been either not publicly available so far or not attaining sufficient performance. We have developed the tools and released the tools to the public, in both software packages and web tools. For development, we utilize state-of-the-art methods and achieved high accuracy. We briefly present the tasks, the methods and the performance of each tool and resource.
This paper presents our work on developing Vietnamese fundamental tools and a resource for analysis. These tools are for speech tagging, diacritics restoration, and orthographical variants dictionary. All of them have been either not publicly available so far or not attaining sufficient performance. We have developed the tools and released them to the public, in both software packages and web tools. For development, we utilize state-of-the-art methods and achieved high accuracy. We briefly present the tasks, the methods and the performance of each tool and resource.
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

<table>
<thead>
<tr>
<th>Isolated Language</th>
<th>Like Chinese. Tôi là sinh viên. (I am a student.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vietnamese Alphabets</td>
<td>Like Chinese pinyin Sequence of syllables sinh nhật (birthday)</td>
</tr>
<tr>
<td>Over 70 Million Speakers</td>
<td>Vietnamese NLP technology helps a lot of people.</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall(%)</th>
<th>Precision(%)</th>
<th>F-Value(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM(Yamcha)</td>
<td>87.83</td>
<td>89.24</td>
<td>88.53</td>
</tr>
<tr>
<td>CRF(CRF++)</td>
<td>89.10</td>
<td>90.10</td>
<td>89.60</td>
</tr>
</tbody>
</table>

SVM  https://github.com/kanjirz50/viet-morphological-analysis-svm
CRF  https://github.com/kanjirz50/viet-morphological-analysis-crf
In some Vietnamese text, diacritics are dropped.

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

83% accuracy.

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

Diacritics mark causes orthographical variation.

<table>
<thead>
<tr>
<th>Syllable 1</th>
<th>Frequency 1</th>
<th>Syllable 2</th>
<th>Frequency 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>thûy</td>
<td>14,280</td>
<td>thuý</td>
<td>2,307</td>
</tr>
<tr>
<td>thûy</td>
<td>6,047</td>
<td>thuý</td>
<td>912</td>
</tr>
<tr>
<td>khoé</td>
<td>1031</td>
<td>khoé</td>
<td>409</td>
</tr>
</tbody>
</table>

These syllables are collected from 1.7 million sentences.

Unicode Problem
• ”ã” 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.