Abstract
Viet Treebank is an annotated corpus newly published in 2009. In this paper, we applied automated methods to detect errors in Viet Treebank based on the concept of equivalence classes proposed by Dickinson. On this basis, we propose an improved method of error detection by transforming the syntax tree based on vertical markovization. Our experimental results on Viet Treebank showed that the scope of error detection extended more than 2 times and improved the precision more than 18.07% in comparison with the base line methods.

Methods
Proposed a limited concept of equivalence classes to detect rare rules, i.e rules may appear very little in corpus.

-Equivalence Class (EC)
Equivalence classes are conducted according to the following steps:
• Remove daughter categories that are always non-predictive to phrase categorization, i.e., always adjuncts such as punctuation and the parenthetical category.
• Group head-equivalent lexical categories, e.g., N (common noun) and NP (proper noun).
• Model adjacent identical elements as a single element, e.g., NP NP becomes NP.
• Whole daughters scoring (WDS)
  • Map the rule to its equivalence class (determine its rule type)
  • Score as formula:
    \[
    \text{score} = \frac{(|E|+1)/2 \sum_{i=1}^{n} (|S_i|)}
    \]
  When \(|E|\) is the number of elements within the equivalence class, \(|S_i|\) is the number of elements of the highly similar equivalence class \(i\) to the rule type

-Bigram Scoring (BGS)
• Reduce the rule under the concept of limited equivalence class. Resulting in a reduced rule (or a rule type).
• Calculate the frequency of each <mother, bigram> pair in the reduced rule: for search occurrence of <mother, bigram> pair in a class, add a score of 1 for that pair.
• Assign score of the lowest-score <mother, bigram> pair as the score of the rule. We do that because we are interested in anomalous sequences.

Results

<table>
<thead>
<tr>
<th>Error set</th>
<th>#Errors detected</th>
<th>#Errors corrected</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC2</td>
<td>427</td>
<td>298</td>
<td>69.79</td>
</tr>
<tr>
<td>EC3</td>
<td>375</td>
<td>264</td>
<td>70.40</td>
</tr>
<tr>
<td>WDS10</td>
<td>315</td>
<td>182</td>
<td>57.78</td>
</tr>
<tr>
<td>WDS5</td>
<td>157</td>
<td>84</td>
<td>53.50</td>
</tr>
<tr>
<td>BGS2</td>
<td>892</td>
<td>597</td>
<td>66.92</td>
</tr>
<tr>
<td>BGS3</td>
<td>457</td>
<td>317</td>
<td>69.37</td>
</tr>
<tr>
<td>Combined</td>
<td>1295</td>
<td>700</td>
<td>54.05</td>
</tr>
</tbody>
</table>

Evaluation of error detection with vertical Markovization

<table>
<thead>
<tr>
<th>Error set</th>
<th># errors detected</th>
<th># errors corrected</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC1</td>
<td>1699</td>
<td>1285</td>
<td>75.63</td>
</tr>
<tr>
<td>WDS5</td>
<td>478</td>
<td>376</td>
<td>78.66</td>
</tr>
<tr>
<td>BGS1</td>
<td>1516</td>
<td>1158</td>
<td>76.38</td>
</tr>
<tr>
<td>Combined</td>
<td>2791</td>
<td>2013</td>
<td>72.12</td>
</tr>
</tbody>
</table>

Conclusion
We found out a significant number of annotation errors in VTB. With these results, we also help linguists that build VTB not only in reducing error detection efforts but also in adjusting annotation guidelines to raise quality of the Vietnamese corpus. The results also showed a significant improvement in error detection scope and precision by transforming syntactic trees to allow additional contextual information about rules to be considered.

References