New Phrase Chunking Algorithm for Myanmar Natural Language Processing

Dr. Myintzu Phyo Aung
Assistant Lecturer
Research and Development Department
University of Computer Studies, Mandalay, Myanmar
Outline

- Abstract
- Introduction
- Proposed Myanmar Phrase Chunking Algorithm
- Myanmar Grammar
- Chunk Markers of Myanmar Language
- Evaluation and Discussion
- Conclusion and Future Work
Abstract

- Chunking is the subdivision of sentences into non-recursive regular syntactical groups.
- The chunker can operate as a preprocessor for Natural Language Processing systems.
- This study aims to propose new phrase chunking algorithm for Myanmar natural language processing.
- The developed new algorithm accepts Myanmar tagged sentence as input and generates chunks as output.
Introduction

- The determination of a correct syntactic tree structure of an input sentence is crucial in Machine Translation.
- Phrase chunking or shallow parsing is a Natural Language Processing (NLP) task which divides a text into phrases, non-overlapping and syntactically correlated parts of words, based on linguistic properties and useful in several tasks.
- Myanmar is a very productive inflectional, agglutinative and grammatical regularity lack language, it is less computerized and lack in computational resources.
The developed new phrase chunking algorithm accepts Myanmar tagged sentence as input and generates chunks as output.

Input Myanmar sentence is split into chunks by using chunk markers such as postpositions, particles and conjunction and define the type of chunks.

The algorithm was evaluated with POS tagged Myanmar sentences based on three measure parameters.
Motivations

- To be successful at any NLP applications of Myanmar language, some amount of functional analysis is necessary to be feasible grammatical relations.

- In the previous Burmese Phrase Segmentation system used the CRF++ tool to identify phrase boundaries with markers such as သည, က, က, ဖ and tested on four types of corpora.

- However, it cannot solve phrases with zero markers and cannot also differentiate homonyms.
Proposed Myanmar Phrase Chunking Algorithm

- Input: POS tagged Myanmar Sentence
- Output: Myanmar Chunks with Chunk Type
- Begin
- Convert POS tagged Myanmar sentence into POS and word string arrays
- While(i<length of POS string array)
  - Read POS[i] of word [i]
  - If POS[i] is chunk marker
    - String array for word is segments before and after word[i]
  - Else
    - Increment i
- End While
- Generate Myanmar chunks with corresponding type of chunks for input sentence
- End
Grammar, the study of rules behind the languages, which does not concern the meaning directly is called the syntax of that language.

Myanmar can be defined as postpositional language because of its use of postposition (Wi Bat),

Myanmar Language Syntax: Subject Object Verb - SOV

English is prepositional language, (syntax: Subject Verb Object - SVO) because of its use of preposition.
Example Myanmar Sentence and Constituent Phrases

<table>
<thead>
<tr>
<th>Sentence</th>
<th>English meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>ေဒ ေ ၚ အ ာ င ဖန္စဖလညက္သည္</td>
<td>Daw Aung San Su Kyi obtained Peace Noble Prize in 1991.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Noun Phrase</th>
<th>Verb Phrase</th>
</tr>
</thead>
</table>
## Noun Chunk Markers of Myanmar Language

<table>
<thead>
<tr>
<th>Myanmar Phrase Chunk Marker</th>
<th>English Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ကတ ၱ ာ း ၀ ◌ ိ ဘတ ္</td>
<td>Nominative</td>
<td>မွ ာ + သ ာ က ာ း ရ + 蒲 း း း</td>
</tr>
<tr>
<td>။ ၂ ိ ။ ိ ိ ိ ိ</td>
<td>Objective</td>
<td>မွ ာ + သ ာ က ာ း ရ</td>
</tr>
<tr>
<td>က ံ ၀ ◌ ိ ဘတ ္</td>
<td>Departure</td>
<td>မ ွ ာ + သ ာ က ာ း ရ</td>
</tr>
<tr>
<td>က ိ ဖg5လ9</td>
<td>Direction</td>
<td>မ ွ ာ</td>
</tr>
<tr>
<td>သ ိ ဖg5လ9 ◌့</td>
<td>Arrival</td>
<td>မ ွ ာ</td>
</tr>
<tr>
<td>၍ ၃ ၃ ၃ ပ ၀ ◌ ိ ဘတ ္</td>
<td>Accusation</td>
<td>မ ွ ာ + သ ာ က ာ း ရ</td>
</tr>
<tr>
<td>၍ ၃ ၃ ၃ ပ ၀ ◌ ိ ဘတ ္</td>
<td>Reason</td>
<td>မ ွ ာ + သ ာ က ာ း ရ</td>
</tr>
<tr>
<td>၍ ၃ ၃ ၃ ပ ၀ ◌ ိ ဘတ ္</td>
<td>Acceptance</td>
<td>မ ွ ာ</td>
</tr>
<tr>
<td>၍ ၃ ၃ ၃ ပ ၀ ◌ ိ ဘတ ္</td>
<td>Time</td>
<td>မ ွ ာ</td>
</tr>
<tr>
<td>၍ ၃ ၃ ၃ ပ ၀ ◌ ိ ဘတ ္</td>
<td>Place</td>
<td>မ ွ ာ</td>
</tr>
<tr>
<td>၍ ၃ ၃ ၃ ပ ၀ ◌ ိ ဘတ ္</td>
<td>Agreement</td>
<td>မ ွ ာ</td>
</tr>
</tbody>
</table>

10
Verb Chunk Marker of Myanmar Language

<table>
<thead>
<tr>
<th>Myanmar Phrase Chunk Marker</th>
<th>English Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>န်မှု့ မဝါ စာ သစ်</td>
<td>Verb Postposition</td>
<td>သည် စာ သစ်</td>
</tr>
<tr>
<td>Myanmar Phrase Chunk Marker</td>
<td>English Meaning</td>
<td>Example</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>----------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>Adjective Phrase Change Particle</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Types of Chunks Identified by the Algorithm

<table>
<thead>
<tr>
<th>No.</th>
<th>Chunk Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>NC</td>
<td>Noun Chunk</td>
</tr>
<tr>
<td>2.</td>
<td>VC</td>
<td>Verb Chunk</td>
</tr>
<tr>
<td>3.</td>
<td>AdjC</td>
<td>Adjective Chunk</td>
</tr>
<tr>
<td>4.</td>
<td>AdvC</td>
<td>Adverb Chunk</td>
</tr>
<tr>
<td>5.</td>
<td>CC</td>
<td>Conjunction Chunk</td>
</tr>
</tbody>
</table>
Results and Discussion

- The algorithm is tested with six types of Myanmar sentences.
- The algorithm uses the POS tags of chunk markers instead of words.
  - It can solve the homonyms.
  - It can not segment the chunks correctly for the sentences with no makers and sentences containing more than one chunk markers continuously.
- The algorithm is tested with 1000 Myanmar POS tagged sentences and results are shown with three measures: Precision, Recall and F-measure.
Cont’d

\[
\text{Precision} = \frac{\text{No. of correctly identified chunks}}{\text{No. of identified chunks}} \times 100\%
\]

\[
\text{Recall} = \frac{\text{No. of correctly identified chunks}}{\text{No. of total chunks in testing sentences}} \times 100\%
\]

\[
F-Measure = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}
\]
### Calculation of Precision, Recall and F-measure

<table>
<thead>
<tr>
<th>Correctly identified chunks</th>
<th>Identified chunks</th>
<th>Total chunks in testing sentences</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>2357</td>
<td>2667</td>
<td>2981</td>
<td>89%</td>
<td>79%</td>
<td>83%</td>
</tr>
</tbody>
</table>
Comparison of percent accuracy for three measures
Conclusion

• Phrase chunking is one of the promising solutions for the ambiguities and produce at certain level of grammatical information.

• The new Myanmar phrase chunking algorithm is implemented. It segments the Myanmar sentence into chunks by using the postpositions, particles and conjunctions.

• The algorithm is tested with POS tagged Myanmar sentences and results are obtained 89% 79% and 84% in Precision, Recall and F-measure respectively.

• This new algorithm is effective and feasible for Myanmar Natural Language Processing applications.
References


Thank You!