Empirical Dependency-Based Head Finalization for Statistical Chinese-, English-, and French-to-Myanmar (Burmese) Machine Translation

Chenchen Ding†, Ye Kyaw Thu‡, Masao Utiyama‡, Andrew Finch‡, Eiichiro Sumita‡

† Department of Computer Science, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573, Japan

‡ National Institute of Information and Communications Technology
3-5 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0289, Japan
Normalized Kendall’s Tau

- Often used to measure differences in word order between languages
- Related to the number of crossing word alignments
- Calculated from word aligned data
- In the range from 0 to 1
  - 0 indicating identical word order
## Kendall’s Tau for Myanmar

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Normalized Kendall’s Tau</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-Myanmar</td>
<td>0.538</td>
</tr>
<tr>
<td>French-Myanmar</td>
<td>0.487</td>
</tr>
<tr>
<td>Chinese-Myanmar</td>
<td>0.315</td>
</tr>
<tr>
<td>Korean-Myanmar</td>
<td>0.156</td>
</tr>
<tr>
<td>Japanese Myanmar</td>
<td>0.123</td>
</tr>
</tbody>
</table>
Importance of re-ordering

![Graph showing the correlation between BLEU and Normalized Kendall's tau. The correlation is 0.8.](image-url)

Correlation = 0.8
Pre-ordering

• Long distance word re-ordering is a problem for SMT
• Pre-ordering approaches have been successful in SVO-to-SOV translation
  • Re-order the source in a pre-processing step
  • Use re-ordering heuristics in combination with a high-precision parser
• Efficient
Exploiting the Head Final Property

• Languages such as Japanese, have the property that the head word typically follows its dependent words

• Parse the source with a head-driven phrase structure grammar (HPSG)
  • Pre-order with rules operating on the parse tree

• English-to-Japanese : [Isozaki+ 2012]
• Chinese-to-Japanese : [Han+ 2012]
Other approaches

• HPSG parsers not available for all languages
  • [Xu+ 2009] proposed using a dependency parser
• Statistical approaches are also possible, e.g. LADER (Neubig+ 2012)
• Myanmar (Burmese) is a typical SOV language
  → Head-finalization should work
Myanmar Language

• SOV language
  ‣ Consistently head-final
  ‣ Similar to Japanese and Korean
  ‣ Function morphemes succeed content morphemes
  ‣ Unlike Japanese and Korean, Myanmar is analytic (morphemes are non-inflected syllables)
Pre-ordering for English-, Chinese-, French-to-Myanmar SMT

- Myanmar (Burmese) Language
  - Similar syntax to Japanese/Korean
    → Transfer techniques used for JA/KO to MY

(Content morphemes in black, function morphemes in grey)
Our Head-Finalization

• Dependency-based head-finalization
  – A combination of [Isozaki+ 2012] and [Xu+ 2009]

• Available for more source languages

• Just move the head after modifiers
  – Simple
  – With several exceptions
    → Examples
Head finalization for Myanmar

Our rules follow 3 basic principles:

• Do not break a coordination structure
  • *English*: conj, cc

• Do not reorder across punctuation
  • *English*: punct

• Auxiliary verbs are placed after their head verb
  • *English*: aux, auxpass, cop
Pre-reordering Examples

- I’ll have a slice of pizza with pepperoni and mushroom.
- Pepperoni and mushroom with pizza of a slice have ‘ll.

- First dial zero, then dial the number you would like to call.
- First zero dial, then you call to like would the number dial.
Myanmar Oriented Process

- In [Isozaki+, 2012] topic, nominative, and accusative markers are inserted on the source side
  - In Myanmar these are typically omitted, unless there is ambiguity
- We handle negation in Myanmar
  - Negation prefix and suffix (“ma ... buu” like “ne ... pas” in French)
  - Place negation word immediately before verb, and “neg” maker immediately after
- Source side articles are deleted
Experimental Methodology

• Chinese-, English-, French-to-Myanmar
  – On BTEC corpus
    • Train: 155,121 sentence pairs
    • Dev.: 5,000 sentence pairs
    • Test: 2,000 sentence pairs

• For dependency parsing
  – Chinese: Stanford parser
  – English: Stanford parser
  – French: Stanford tagger (CC set) + MALT parser

• Statistical approach for comparison
  – LADER
Myanmar Segmentation

- 2 approaches
  - Syllable segmentation (Fr-My)
  - Maximum matching (Ch-My, En-My)
Training LADER

- 1000 sentences sampled randomly
  - Used automatic alignment since no manually-aligned data was available
- Using larger training data set sizes with automatic alignment gave only a small improvement in the original work [Neubig+ 2012].
- Long training times.
Evaluation

• Evaluated using BLEU (default MOSES)
• Also used RIBES
  • For distant language pairs, BLEU has been shown to correlate poorly with human judgements [Goto+ 2011]
  • RIBES was developed specifically to evaluate distant language pairs
• Results had similar characteristics with both metrics
Results of Chinese-to-Myanmar

- Average Kendall’s $\tau$
  - Baseline : 0.31
  - LADER : 0.20
  - Head Final : 0.17
- Maximum matching segmentation
Results of English-to-Myanmar

- Average Kendall’s τ
  - Baseline : 0.47
  - LADER : 0.21
  - Head Final : 0.21
- Maximum matching segmentation

![Graph showing Test Set BLEU vs Distortion-Limit](image)
Results of French-to-Myanmar

- Average Kendall’s $\tau$
  - Baseline : 0.46
  - LADER : 0.24
  - Head Final : 0.24

- Maximum matching segmentation
Summary

• Head-final approach works on distant languages to Myanmar SMT
  – Can use dependency structure
  – Simple set of rules

• Future work
  – Experiment on larger corpora
  – Experiment on longer sentences
Thank you very much for listening!