Overview
Dependency-based pre-ordering for Zh-Ja MT
- Patent-adapted in-house dependency parser
- Two-types of pre-ordering:
  * Rule-based, Head Final Chinese (Han+ 2012)
  * Data-driven, Learning to Rank (Yang+ 2012)
- Rule-based system is better, comparable to T2S

Syntactic Analysis
[Word segmentation & POS tagging]
- Joint sequential labeling (Suzuki+ 2012)
[Dependency parsing (untyped)]
- Second-order graph-based parsing
[Semi-supervised learning] (Suzuki+ 2009)
- Labeled: 31K sents. (news), 35K sents. (patents)
- Unlabeled: 9GB (news), 100GB (patents)

Table 1: Performance in Chinese syntactic analysis

<table>
<thead>
<tr>
<th></th>
<th>Word seg.</th>
<th>POS</th>
<th>Dep.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>F0 / UAS</td>
<td>0.927</td>
<td>0.855</td>
</tr>
</tbody>
</table>

References:
Han, Dan et al., Head Finalization Reordering for Chinese-to-Japanese MT, Proc. SSST-6 (2012)
Suzuki, Jun et al., \(\ldots\) [in Japanese]

Rule-based pre-ordering
Base rule: Moving a head word after its modifiers
Exceptions (placed after their head words):
- AS (aspect particle), SP (sentence-final particle)
- PU (punctuation), CC (coordinating conjunction)
- IJ (interjection), “不” (negation), “等” (“etc.”)

Pros: stability, domain independence (?)
Cons: effort for rule management

Data-driven pre-ordering
Reordering by reranking a head & its modifiers (Yang+ 2012)
- Implemented with Ranking SVM
  * Features: - surface/POS (head & modifier)
    - modifier surface/POS (h & m)
    - span surfaces/POSs (modifier)
    - relative position (h & m)
  * Reordering oracles are determined by maximizing Kendall’s tau criterion (Hoshino+ 2015)

Pros: no special effort, target adaptability
Cons: instability, noisy auto. word alignment

SMT setup
Standard Moses Phrase-based MT
- MGIZA word alignment, g-d-f-a symal
- Kneser-Ney phrase-table score smoothing
- Word 5-gram LM with Kneser-Ney smoothing
- Distortion limit: 9 (chosen over 0,3,6,9)
- Weights chosen over 5 indep. MERT runs

Comparable to T2S baseline
Rule-based is better than data-driven

Results
Table 2: Official evaluation results

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>RIBES</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL PBMT</td>
<td>n/a</td>
<td>0.781</td>
<td>0.382</td>
</tr>
<tr>
<td>BL T2S</td>
<td>20.75</td>
<td>0.814</td>
<td>0.394</td>
</tr>
<tr>
<td>Rule-based</td>
<td>16.25</td>
<td>0.822</td>
<td>0.406</td>
</tr>
<tr>
<td>Data-driven</td>
<td>8.00</td>
<td>0.812</td>
<td>0.399</td>
</tr>
</tbody>
</table>

Conclusion
Pre-ordering is a deterministic approx. of T2S
--- good in efficiency with some loss in accuracy
> forest-based pre-ordering, pre-ordering lattice
Rule-based pre-ordering works robustly
--- due to head-final nature in Japanese
Data-driven pre-ordering is still challenging...
--- difficulty in word alignment, non-parallelism
--- constituent or dependency structures?
Remained patent MT issues:
- Context awareness (consistency)
- Domain awareness (lexical choice)