Evaluating Neural Machine Translation in English-Japanese Task
(TEAM ID: WEBLIO MT)

Zhongyuan Zhu  
@raphaelshu
Empirically evaluate various models in EJ task

- Two network architectures
- Three recurrent units
  - LSTM, GRU, IRNN
- Two kinds of training data
  - naturally-ordered, pre-reordered
Results: perplexities
# Results: evaluation scores

|                                | BLEU | RIBES | HUMAN | JPO |
|--------------------------------|------|-------|-------|-----|
| Baseline phrase-based SMT      | 29.80| 0.691 |       |     |
| Baseline hierarchical phrase-based SMT | 32.56| 0.746 |       |     |
| Baseline Tree-to-string SMT    | 33.44| 0.758 | 30.00 |     |
| Submitted system 1 (NMT)       | 34.19| 0.802 | 43.50 |     |
| Submitted system 2 (NMT + System combination) | 36.21| 0.809 | 53.75 | 3.81|
| Best competitor 1: NAIST (Travatar System with NeuralMT Reranking) | 38.17| 0.813 | 62.25 | 4.04|
| Best competitor 2: naver (SMT t2s + Spell correction + NMT reranking) | 36.14| 0.803 | 53.25 | 4.00|
Finding & Insights

- Soft-attention models outperform multi-layer encoder-decoder models
- Training models on pre-reordered data hurts the performance
- NMT models tend to make grammatically valid but incomplete translations
Thanks.