Neural Reranking Helps Subjective Quality of Machine Translation: NAIST at WAT 2015
Graham Neubig, Makoto Morishita, Satoshi Nakamura
Nara Institute of Science and Technology (NAIST), Japan

Research Questions
1. Does reranking with neural MT models improve subjective impressions of translation results?
2. If so, what are the qualitative differences between reranked and non-reranked output?
3. How big of an n-best list do we need?

Quantitative Results

Detailed Analysis

#1 Improvement: Phrasal Reordering (+26, -4)
Source: In case 2, reddening, induration, and skin ulcer appeared during chemical therapy for liver metastasis of rectal cancer.
Base: In case 2, occurred during chemotherapy for liver metastasis of rectal cancer, flare, induration, skin ulcer.
Rerank: In case 2, flaring, induration, skin ulcer was produced during the chemotherapy for hepatic metastasis of rectal cancer.

#2 Improvement: Auxiliary Verb Ins./Del. (+15, -0)
Source: これにより得られる支配方程式は壁面乱流のような非線形化にも有用である。
Ref: Governing equation derived by this method is useful for turbulent shear flow like turbulent flow near wall.
Base: The governing equation is obtained by this is also useful for shear flow such as wall turbulence shear flow.
Rerank: The governing equation obtained by this is also useful for shear flow such as wall turbulence.

#3 Improvement: Coordinate Structures (+13, -2)
Source: ラングミュア−ブロジェット法や包接化にも触れた。
Ref: Coordination structures are useful in the stress analysis.
Base: The infrared application measurement using Langmuir-Blodgett method and inclusion compounds are also discussed.
Rerank: The infrared application measurement using Langmuir-Blodgett method and inclusion are also mentioned.

What Didn't Work: Terminology (+2, -4)

#4 Improvement: Verb Agreement (+6, -0)
Source: Infrared ray applied measurement using radiant heat is useful for stress analysis.
Ref: The infrared application measurement using radiant heat is useful in the stress analysis.
Base: The infrared application measurement using radiant heat is useful for stress analysis.
Rerank: The infrared application measurement using radiant heat is useful for stress analysis.

Conclusion
• Reranking with neural MT models leads to subjective improvements in MT quality
• Future work includes comparisons with neural language models or neural MT w/o reranking

Neural Reranking

Experimental Setup
• Data: ASPEC Scientific Abstracts
  - Japanese ↔ English, Chinese
• Baseline: NAIST WAT2014 Tree-to-String System
  - Strong baseline achieving high scores
  - Implemented using Travatar
  (http://phontron.com/travatar)
• Neural MT Model: Attentional model
  - Trained ~500k sent., 256 hidden nodes, 2 model ensemble
  - Use words occurring 3+ times (vocab 50,000~80,000)
  - Trained w/ lamtram
  (http://github.com/neubig/lamtram)
• Automatic Evaluation: BLEU, RIBES
• Manual Evaluation: WAT 2015 HUMAN Score

Quantitative Results

Neural Mount Model

T2S

Rescored/Reranked N-best
1. 彼は寒さを持っており t=-0.5 l=-5.5 nmt=-5.8
2. 彼は風邪を持っている t=-0.9 l=-5.8 nmt=-5.5
3. 彼は風邪を引いた t=-1.5 l=-5.3 nmt=-3.4
4. 彼は風邪がある t=-1.9 l=-5.4 nmt=-5.2

Reranking and N-best Size

Experimental Setup

Type Improved Degraded % Impr.
Reordering 55 9 86%
Deletion 20 10 67%
Insertion 19 2 90%
Substitution 15 11 58%
Conjugation 8 1 89%
Total 117 33 78%

Overall improvements re-confirmed
Particularly reordering, insertion, and conjugation errors

Qualitative Analysis