Neural Reranking Improves Subjective Quality of Machine Translation: NAIST at WAT 2015

Graham Neubig, Makoto Morishita, Satoshi Nakamura
Nara Institute of Science and Technology (NAIST)
2015-10-16
Statistical Translation Frameworks

Symbolic Models

**Phrase-based MT** [Koehn+ 03]

- he has a cold
- 彼 は 風邪 を 引いている
- 彼 は 風邪 を 引いている

**Tree-to-String MT** [Liu+ 06]

- 彼 は 風邪 を 引いている

Continuous-space (Neural) Models

**Encoder-Decoder** [Sutskever+ 14]

- he has a cold <s> 彼 は 風邪 を 引いて いる

**Attentional** [Bahdanau+ 15]

- P(e_i|F,e_1,...,e_{i-1})
Relative Merits/Demerits

- **Symbolic Models**
  - Inner workings well understood
  - Better at translating low-frequency words

- **Continuous-space Models**
  - Easier to implement
  - Produce more fluent output
  - Probabilistic model – can score output of other systems!
Reranking with Neural MT Models

<table>
<thead>
<tr>
<th>Input</th>
<th>N-best w/MT Features</th>
<th>Neural Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>he has a cold</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|                       | t=-0.5 l=-5.6 | -6.1          | nmt=-5.8         |
| 1. 彼は寒さを持っている |                |                |                  |
| 2. 彼は風邪を持っている | t=-0.9 l=-5.8 | -6.7          | nmt=-5.5         |
| 3. 彼は風邪を引いた   | t=-1.5 l=-5.3 | -6.8          | nmt=-3.4         |
| 4. 彼は風邪がある    | t=-1.9 l=-5.4 | -7.3          | nmt=-5.2         |

Rescored/Reranked N-best

|                       | t=-1.5 l=-5.3 | nmt=-3.4 | -9.2          |
| 1. 彼は風邪を引いた   |                |          |                |
| 2. 彼は寒さを持っている | t=-0.5 l=-5.6 | nmt=-5.8 | -10.9         |
| 3. 彼は風邪を持っている | t=-0.9 l=-5.8 | nmt=-5.5 | -11.2         |
| 4. 彼は風邪がある    | t=-1.9 l=-5.4 | nmt=-5.2 | -12.5         |
What Do We Know About Reranking?

- Reranking greatly improves BLEU score, even over strong baseline systems:

<table>
<thead>
<tr>
<th></th>
<th>Sutskever+ 2014</th>
<th>Alkhouli+ 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>en-fr</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>33.3</td>
<td>30.6</td>
</tr>
<tr>
<td>Rerank</td>
<td>36.5</td>
<td>32.3</td>
</tr>
<tr>
<td><strong>de-en</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>26.4</td>
<td></td>
</tr>
<tr>
<td>Reranked</td>
<td>27.0</td>
<td></td>
</tr>
</tbody>
</table>
What **Don't** We Know About Reranking?

- Does reranking improve subjective impressions of results?
- What are the qualitative differences before/after reranking with neural MT models?
Experiments
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
Neural Reranking Improves Subjective Quality of Machine Translation

Results

Confirm what we know: Neural reranking helps automatic evaluation.

Show what we didn't know: Also help manual evaluation.
What is Getting Better?

- Perform detailed categorization of the changes in Japanese-English results:
  1. Is the sentence better/worse after ranking?
  2. What is the main error corrected: insertion, deletion, substitution, reordering, or conjugation?
  3. What is the detailed subcategory?
Main Types of Errors Corrected/Caused

<table>
<thead>
<tr>
<th>Type</th>
<th>Improved</th>
<th>Degraded</th>
<th>% Impr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reordering</td>
<td>55</td>
<td>9</td>
<td>86%</td>
</tr>
<tr>
<td>Deletion</td>
<td>20</td>
<td>10</td>
<td>67%</td>
</tr>
<tr>
<td>Insertion</td>
<td>19</td>
<td>2</td>
<td>90%</td>
</tr>
<tr>
<td>Substitution</td>
<td>15</td>
<td>11</td>
<td>58%</td>
</tr>
<tr>
<td>Conjugation</td>
<td>8</td>
<td>1</td>
<td>89%</td>
</tr>
<tr>
<td>Total</td>
<td>117</td>
<td>33</td>
<td>78%</td>
</tr>
</tbody>
</table>

Overall improvements re-confirmed

In particular fixing reordering, insertion, and conjugation errors
#1 Detailed Improvement Category: Phrasal Reordering (+26, -4)

**Source**

症例2においては、直腸がんの肝転移に対する化学療法中に、発赤、硬結、皮膚潰ようを生じた。

**Ref**

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, the flare, induration, skin ulcer was produced during the chemotherapy for hepatic metastasis of rectal cancer.
#2 Detailed Improvement Category: 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せん断乱流にも有用である。

**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 such as wall turbulence shear flow.

**Rerank**

The governing equation obtained by this is also useful for shear flow such as wall turbulence.
#3 Detailed Improvement Category: Coordinate Structures (+13, -2)

**Source**
レーザー加工は高密度光束による局所的な加熱とアブレーションにより行う。

**Ref**
Laser work is done by local heating and ablation with high density light flux.

**Base**
The laser processing is carried out by local heating by high-density luminous flux and ablation.

**Rerank**
The laser processing is carried out by local heating and ablation by high-density flux.
#4 Detailed Improvement Category: Verb Agreement (+6, 0)

Source
ラングミュア-ブロジェット法や包接化にも触れた。

Ref
Langmuir-Blodgett method and inclusion compounds are mentioned.

Base
Langmuir-Blodgett method and inclusion is also discussed.

Rerank
Langmuir-Blodgett method and inclusion are also mentioned.
What Wasn't Helped: Terminology (+2, -4)

Source
放射熱を利用する赤外線応用計測が応力解析に役立っている

Ref
Infrared ray applied measurement using radiant heat is useful for stress analysis.

Base
The infrared application measurement using radiant heat is useful in the stress analysis.

Rerank
Infrared ray application measurement using radiation heat is useful for stress analysis.
Conclusion
What Do We Know Now?

- Neural reranking improves subjective quality of machine translation output.

- Main gains are from grammatical factors, and not lexical selection.
What Do We Still Not Know Yet?

• How do neural translation models compare with neural language models?

• How does reranking compare with pure neural MT?
Thank You!