Maximum Rank Correlation Training for Statistical Machine Translation

Daqi Zheng*, Yifan He**, Yang Liu*, Qun Liu* **
*ICT, Chinese Academy of Sciences
**CNGL, Dublin City University
Originally published at MT Summit XIII 2011

Presented by Qun Liu
MONOMT Workshop
Attached with AMTA, San Diego US, 1st Nov 2012
Outline

Introduction

Motivation

Maximum Rank Correlation Training

Experiments and Results

Result Analysis

Conclusion
Introduction

• MERT: Minimal Error Rate Training
  – N-best candidates are given by the decoder for each sentences
  – Tune the parameter $\lambda$ to make the best candidate (with highest BLEU score) to have the highest model score
MERT

- Not good for rich features (>20)
- Not stable for local extremums
- Not generalizable across domains
Alternative solution: Min-Risk

[Li & Eisner EMNLP2009]

• Define the Risk as:

\[ R = - \sum_{i} p(c_i) \text{SBLEU}(c_i) \]

while the Posterior Probability can be defined using model score, for example:

\[ p(c_i) = \frac{\exp(\gamma \cdot \text{score}(c))}{\sum_i \exp(\gamma \cdot \text{score}(c_i))} \]

• Tune the parameter \( \lambda \) to minimize the Risk
Alternative solution: MIRA

[Chiang et al. EMNLP2008]

- Select a **Positive Set** of candidates with high BLEU scores
- Select a **Negative Set** of candidates with low BLEU score
- Tune the parameter $\lambda$ to **maximize the difference (margin)** of the model score between Negative Set and that of Positive Set.
Example

![Graph showing model score against parameter \( \lambda \). The graph has two lines: one for POS and one for NEG. The lines intersect at different values of \( \lambda \) and show the margin between POS and NEG. The labels MERT, Min-Risk, and MIRA are indicated on the x-axis.]
Our Motivation: Max Rank Correlation

• We would like to choose the $\lambda$ which maximize the correlation between the ranking of the candidates according the model scores and that according to the BLEU scores
Motivation

• For example: 8 candidates
  – BLEU score ranking: 1 2 3 4 5 6 7 8
  – Model score ranking with $\lambda_1$: 1 8 7 6 5 4 3 2
  – Model score ranking with $\lambda_2$: 2 1 3 4 5 6 7 8

• For MERT, $\lambda_1$ will be chosen
• We would like to choose $\lambda_2$
<table>
<thead>
<tr>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
</tr>
<tr>
<td>Motivation</td>
</tr>
<tr>
<td>Maximum Rank Correlation Training</td>
</tr>
<tr>
<td>Experiments and Results</td>
</tr>
<tr>
<td>Result Analysis</td>
</tr>
<tr>
<td>Conclusion</td>
</tr>
</tbody>
</table>
Maximum Rank Correlation (MRC)

\[ \hat{\lambda} = \arg \max_{\lambda} \left( \sum_{i=1}^{M} w_i \cdot Corr_i(\lambda) \right) \]

\[ Corr_i(\lambda) = Corr(\Phi_1^N(\lambda), SBLEU(e_1^N)) \]

Spearman Rank Correlation Coefficient

\[ \rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \]

\[ w_i = \text{length}(f_i) / \sum_{i=1}^{M} \text{length}(f_i) \]
Example
Combination of MER and MRC

\[ \hat{\lambda} = \operatorname{arg\,max}_{\lambda} \left( \alpha \cdot \sum_{i=1}^{M} \text{Corr}_i(\lambda) + (1 - \alpha) \cdot \text{BLEU}(\lambda) \right) \]
Multi-objective Optimization

• We use multi-objective evolutionary algorithm (MOEA) (Fonseca et al., 1993) for training.

• We choose an effective MOEA tool: NSGA-II in our experiments.
Experiment Settings

• Data
  – French-English WMT08 shared translation task
  – Training data: Europarl v3b release
  – Language model: English part of monolingual language model
  – training data
  – Tuning set: dev2006

• System
  – Machine Translation: Moses Suite
  – Spearman Rank Correlation Coefficient: goose
  – Multi-Objective Optimization: NSGA-II
Generic Algorithm Settings

• First Generation
  – 10 individuals from MERT training
  – 390 individuals randomly generated

• Evolution
  – 100 generations
  – 400 individuals for each generation
Experiment Process

Baseline

Reranking

Decoding
Baseline, Reranking and Decoding

1. Previous Round MERT
2. Nbest
3. Parameter
4. Last Round MERT
5. MERT's Best Parameter
6. Decoder
7. Test Set's Nbest
8. Pick the 1best
9. Baseline
10. NSGA-II
11. MRC's Best Parameter
12. Reranker
13. Reranking's result
14. Decoding's result
Baseline, Reranking and Decoding

1. Previous Round MERT
   - Nbest
   - Last Round MERT
   - MERT’s Best Parameter

2. NSGA-II
   - MRC’s Best Parameter
   - Decoder
   - Test Set’s Nbest
   - Pick the 1best
   - Baseline
   - Reranker
   - Reranking’s result

3. Decoder
   - MRC’s Nbest
   - Pick the 1best
   - Decoding’s result
Results

• Best $\alpha$ on development set
• Results via different $\alpha$ on test set
• Improvement of reranking on each MERT tuning run
• Improvement of reranking on different genetic algorithm settings
• Time cost
Result of Best $\alpha$ on Dev Set

![Graph showing BLEU scores for reranking, decoding, and baseline](image-url)
Results

- Best $\alpha$ on development set
- **Results via different $\alpha$ on test set**
- Improvement of reranking on each MERT tuning run
- Improvement of reranking on different genetic algorithm settings
- Time cost
Result via Different $\alpha$

In-Domain

Out-of-Domain
Result via Different $\alpha$

In-Domain

**test2006**

**test2007**

**test2008**

Out-of-Domain

**newstest2008**

**newstest2009**

**newstest2010**
Results

- Best $\alpha$ on development set
- Results via different $\alpha$ on test set
- **Improvement of reranking on each MERT tuning run**
- Improvement of reranking on different genetic algorithm settings
- Time cost
Baseline, Reranking and Decoding

- Previous Round MERT
  - Nbest
    - Last Round MERT
      - MERT's Best Parameter
  - Parameter
    - NSGA-II
      - MRC's Best Parameter
        - Decoder
          - Test Set's Nbest
            - Pick the 1best
              - Baseline
          - Reranker
            - Reranking's result

Dev. Set

Test Set

Baseline

Reranking
Improvement of Reranking on Each Run

In-Domain

Out-of-Domain
Improvement of Reranking on Each Run

In-Domain

Out-of-Domain
Improvement of Reranking on Each Run

In-Domain

Out-of-Domain
Results

• Best $\alpha$ on development set
• Results via different $\alpha$ on test set
• Improvement of reranking on each MERT tuning run
• Improvement of reranking on different genetic algorithm settings
• Time cost
Result Summary

![Graph showing BLEU variance for different GA settings and datasets. The x-axis represents different GA settings, and the y-axis shows the BLEU variance. The graph includes bars for various datasets like test2005, test2006, etc., with error bars indicating variance.](image-url)
Results

• Best $\alpha$ on development set
• Results via different $\alpha$ on test set
• Improvement of reranking on each MERT tuning run
• Improvement of reranking on different genetic algorithm settings
• **Time cost**
## Time Cost

<table>
<thead>
<tr>
<th>Index of Exps</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Iteration</td>
<td>11</td>
<td>15</td>
<td>15</td>
<td>12</td>
<td>9</td>
<td>10</td>
<td>15</td>
<td>12</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Last Mert</td>
<td>26</td>
<td>22</td>
<td>28</td>
<td>23</td>
<td>17</td>
<td>13</td>
<td>23</td>
<td>12</td>
<td>18</td>
<td>26</td>
</tr>
<tr>
<td>Tuning</td>
<td>1222</td>
<td>1766</td>
<td>1657</td>
<td>1329</td>
<td>1047</td>
<td>1113</td>
<td>1694</td>
<td>1363</td>
<td>1613</td>
<td>1741</td>
</tr>
<tr>
<td>$len^1 &amp;(400 \cdot 100)$</td>
<td>780</td>
<td>615</td>
<td>782</td>
<td>793</td>
<td>570</td>
<td>448</td>
<td>755</td>
<td>499</td>
<td>606</td>
<td>793</td>
</tr>
<tr>
<td>$len^2 &amp;(400 \cdot 100)$</td>
<td>768</td>
<td>624</td>
<td>769</td>
<td>778</td>
<td>569</td>
<td>431</td>
<td>718</td>
<td>499</td>
<td>607</td>
<td>741</td>
</tr>
<tr>
<td>$len^2 &amp;(100 \cdot 50)$</td>
<td>95</td>
<td>76</td>
<td>96</td>
<td>97</td>
<td>70</td>
<td>55</td>
<td>89</td>
<td>63</td>
<td>75</td>
<td>96</td>
</tr>
</tbody>
</table>

- 10 experiment’s Running Time: in 100 seconds. Compare the GA with the total tuning time, and consider it need only run once at the tuning phase, the computation cost is affordable.
<table>
<thead>
<tr>
<th>Index of Exps</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Iteration</td>
<td>11</td>
<td>15</td>
<td>15</td>
<td>12</td>
<td>9</td>
<td>10</td>
<td>15</td>
<td>12</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Last Mert Tuning</td>
<td>26</td>
<td>22</td>
<td>28</td>
<td>23</td>
<td>17</td>
<td>13</td>
<td>23</td>
<td>12</td>
<td>18</td>
<td>26</td>
</tr>
<tr>
<td>len₁ &amp; (400 · 100)</td>
<td>1222</td>
<td>1766</td>
<td>1657</td>
<td>1329</td>
<td>1047</td>
<td>1113</td>
<td>1694</td>
<td>1363</td>
<td>1613</td>
<td>1741</td>
</tr>
<tr>
<td>len₂ &amp; (400 · 100)</td>
<td>780</td>
<td>615</td>
<td>782</td>
<td>793</td>
<td>570</td>
<td>448</td>
<td>755</td>
<td>499</td>
<td>606</td>
<td>793</td>
</tr>
<tr>
<td>len₂ &amp; (100 · 50)</td>
<td>768</td>
<td>624</td>
<td>769</td>
<td>778</td>
<td>569</td>
<td>431</td>
<td>718</td>
<td>499</td>
<td>607</td>
<td>741</td>
</tr>
</tbody>
</table>

- 10 experiment’s Running Time: in 100 seconds. Compare the GA with the total tuning time, and consider it need only run once at the tuning phase, the computation cost is affordable.
### Time Cost

<table>
<thead>
<tr>
<th>Index of Exps</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Iteration</td>
<td>11</td>
<td>15</td>
<td>15</td>
<td>12</td>
<td>9</td>
<td>10</td>
<td>15</td>
<td>12</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Last Mert</td>
<td>26</td>
<td>22</td>
<td>28</td>
<td>23</td>
<td>17</td>
<td>13</td>
<td>23</td>
<td>12</td>
<td>18</td>
<td>26</td>
</tr>
<tr>
<td>Tuning</td>
<td>1222</td>
<td>1766</td>
<td>1657</td>
<td>1329</td>
<td>1047</td>
<td>1113</td>
<td>1694</td>
<td>1363</td>
<td>1613</td>
<td>1741</td>
</tr>
<tr>
<td>(l_{en}^1&amp;(400 \cdot 100))</td>
<td>780</td>
<td>615</td>
<td>782</td>
<td>793</td>
<td>570</td>
<td>448</td>
<td>755</td>
<td>499</td>
<td>606</td>
<td>793</td>
</tr>
<tr>
<td>(l_{en}^2&amp;(400 \cdot 100))</td>
<td>768</td>
<td>624</td>
<td>769</td>
<td>778</td>
<td>569</td>
<td>431</td>
<td>718</td>
<td>499</td>
<td>607</td>
<td>741</td>
</tr>
<tr>
<td>(l_{en}^2&amp;(100 \cdot 50))</td>
<td>95</td>
<td>76</td>
<td>96</td>
<td>97</td>
<td>70</td>
<td>55</td>
<td>89</td>
<td>63</td>
<td>75</td>
<td>96</td>
</tr>
</tbody>
</table>

- 10 experiment’s Running Time: in 100 seconds. Compare the GA with the total tuning time, and consider it need only run once at the tuning phase, the computation cost is affordable.
Result Analysis

• MERT+MRCT outperforms MERT both for in-domain and out-of-domain test data

• Why?
BLEU Score vs. Model Score

![Graph showing BLEU Score vs. Model Score]
MER Training

• MER Training tries to make the right most dot at the highest position
• MER Training does not care if the rest of the line is monotone
MER Training

BLEU Score vs. Model Score

![Graph showing BLEU Score vs. Model Score with vertical lines at C₁, C₂, ..., Cₙ and a line labeled λ₁.](image-url)
Max-Margin Training (MIRA)

- Max-Margin Training focuses the positive candidates and the negative candidates
- Max-Margin Training tries to maximize the margin of the model scores between positive candidates and the negative candidates
- Max-Margin Training does not care about the model scores of the medial candidates
Max-Margin Training (MIRA)

BLEU Score vs. Model Score

Model Score

BLEU Score

Positive

Negative

Margin

λ1
Min-Risk Training

- Min-Risk training tends to maximize the model score of the candidate with the highest BLEU score, while minimize the model scores of all other candidates.
- Min-Risk does not care if the line is monotone or not.
Min-Risk Training

BLEU Score vs. Model Score

Model Score vs. BLEU Score

λ1
MRC Training

• MRC Training tries to make the whole line most looks monotone
• MRC Training does not ensure the right most dot be the highest one
MERT + MRCT

- MRCT may be regarded as a regularization for MERT
  - There are many possible choices which satisfy the MER criteria, while some of these choices are severely non-monotone
  - The MRCT helps to choose the parameter which most looks monotone, while satisfy the MER criteria

- That’s the reason why:
  \[ \text{MERT} + \text{MRCT} > \text{MERT} \]
Future Question

• Why the improvements of MERT+MRCT on in-domain test data is much larger than that on out-of-domain test data?
Answer (1/4)

• From the in-domain training data, we obtain both in-domain knowledge and general-domain knowledge.
• In the decoding process, in-domain knowledge and general-domain knowledge are in competition.
Answer (2/4)

• In the n-best list, some candidates are translated using more in-domain knowledge, while some are using more general-domain knowledge.

• The candidates translated using more in-domain knowledge usually get higher BLEU score because the references is given by in-domain development set.
MER Training

BLEU Score vs. Model Score

Model Score

BLEU Score

general-domain

in-domain

$\lambda_1$
Answer (3/4)

• We may find that the in-domain part of the MERT line is basically monotone, while the general-domain part is not.

• But the MRCT line is almost monotone for all parts.
Model Space

- Consider a space consist of all models, where each model is a dot in the space.
- The models perform well in general domain are distributed different with those perform well in specific domains.
Model Space

General-Domain Performance

In-Domain Performance

Out-of-Domain Performance
MERT

In-Domain Performance

General-Domain Performance

Out-of-Domain Performance
Answer (4/4)

- We can see that the model trained using MERT+MRCT will gain better performance on general-domain test data, as well as on out-of-domain test data, even if we do not use the out-of-domain data for training.
Conclusion

• We propose a Maximum Rank Correlation Training approach for parameter tuning for SMT
• We using a multi-objection generative algorithm for parameter tuning
• MRCT + MERT performs a little bit better than MERT for in-domain test data, but much more better for out-of-domain test data
• The time cost of MRCT training is acceptable
• We give an reasonable explanation to the results
THANKS!

Q & A