ListNet-based MT Rescoring

Jan Niehues, Quoc Khanh Do, Alexandre Allauzen and Alex Waibel
Motivation

- Log-linear model is widely used in SMT
  - Use during decoding
  - Use in MT rescoring
- MT Rescoring
  - Easy and efficient way to integrate of complex models
- Machine learning view
  - Ranking problem
  - Promising approach: ListNet algorithm
- Apply ListNet algorithm to SMT
Related Work

Optimization in Machine translation

- Minimum Error Rate Training (MERT) (Och, 2003)
  - Standard in most machine translation systems
- MIRA (Watanabe et al., 2007; Chiang et al., 2008)
- PRO (Hopkins and May, 2011)
- Expected BLEU (Rosti et al, 2011; He and Deng, 2012)

Ranking in machine learning

- ListNet algorithm (Cao et al., 2007)
- Overview over different ranking algorithms (Chen et al., 2009)
Overview

- Motivation
- ListNet Algorithm
- MT Rescoring
  - MT specific problems
- Evaluation
  - WMT
  - IWSLT - TED
ListNet - Ranking

- Input:
  - List
  - Model score
  - Metric for reference ranking

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Model</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7.4</td>
<td>24.4</td>
</tr>
<tr>
<td>B</td>
<td>7.8</td>
<td>24.2</td>
</tr>
<tr>
<td>C</td>
<td>7.2</td>
<td>24.5</td>
</tr>
<tr>
<td>D</td>
<td>7.1</td>
<td>24.1</td>
</tr>
</tbody>
</table>
ListNet - Ranking

- Input:
  - List
  - Model score
  - Metric for reference ranking

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Model</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>7.8</td>
<td>24.2</td>
</tr>
<tr>
<td>A</td>
<td>7.4</td>
<td>24.4</td>
</tr>
<tr>
<td>C</td>
<td>7.2</td>
<td>24.5</td>
</tr>
<tr>
<td>D</td>
<td>7.1</td>
<td>24.1</td>
</tr>
</tbody>
</table>

According to the model
ListNet - Ranking

- Input:
  - List
  - Model score
  - Metric for reference ranking

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Model</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.2</td>
<td>24.5</td>
</tr>
<tr>
<td>A</td>
<td>7.4</td>
<td>24.4</td>
</tr>
<tr>
<td>B</td>
<td>7.8</td>
<td>24.2</td>
</tr>
<tr>
<td>D</td>
<td>7.1</td>
<td>24.1</td>
</tr>
</tbody>
</table>

According to the metric
ListNet - Ranking

- **Input:**
  - List
  - Model score
  - Metric for reference ranking

### Input Metrics

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Model</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>7.8</td>
<td>24.2</td>
</tr>
<tr>
<td>A</td>
<td>7.4</td>
<td>24.4</td>
</tr>
<tr>
<td>C</td>
<td>7.2</td>
<td>24.5</td>
</tr>
<tr>
<td>D</td>
<td>7.1</td>
<td>24.1</td>
</tr>
</tbody>
</table>

### Aim:

Learn a model to rank like the metric
ListNet - Idea

- Define a probability distribution over possible rankings
- Learn model that produces a distribution similar to the one defined by the metric
- Problem: large number of possible rankings
- Define a probability distribution associated to the model ranking based on first ranked object

\[
P_s(j) = \frac{\exp(s_j)}{\sum_{k=1}^{n} \exp(s_k)}
\]  

(1)
ListNet - Distribution

- Minimize cross-entropy difference
Overview

- Motivation
- ListNet Algorithm
- MT Rescoring
  - MT specific problems
- Evaluation
  - WMT
  - IWSLT - TED
MT Rescoring

- Use ListNet to rescore N-Best list
  - Train log-linear model
- Input:
  - N-Best list
  - Additional features
- Learn new weights for log-linear model
**Model**

- Define probability distribution associated to the model ranking

\[ P_s(j) = \frac{\exp(s_j)}{\sum_{k=1}^{n} \exp(s_k)} \]  

(2)

- Problem:
  - Many scores are small probabilities
  - Log-probabilities are very small negative values
  - \( \exp(s) \) calculation may be erroneous

- Feature normalization:
  - Linear transform all features to the range \([-1, 1]\)

- Score normalization:
  - Linear transform the final score of the model to the range \([-r, r]\)
Define probability distribution associated to the reference ranking
- Reference ranking for every sentence needed
- Ranking induced by MT metric
- Sentence-wise MT metric
  - Metric: BLEU+1 (Liang et al. 2006)
  - Smoothed version of BLEU score

\[
P_{y^{(i)}(x_j^{(i)})} = \frac{\exp(\text{BLEU}(x_j^{(i)}))}{\sum_{j' = 1}^{n^i} \exp(\text{BLEU}(x_{j'}^{(i)}))}
\]
Minimize cross-entropy difference between model-based and BLEU+1-based probability distribution

- Use ListNet algorithm to calculate derivation

Stochastic gradient descent

- 100,000 batches
- Batch size of 10
Overview

- Motivation
- ListNet Algorithm
- MT Rescoring
  - MT specific problems
- Evaluation
  - WMT
  - IWSLT - TED
Evaluation

- WMT 2015 EN-DE
  - PBMT System
  - Additional features based on neural network translation models
- WMT 2015 DE-EN
  - PBMT System
  - Additional features using RBM-based translation models and source DWL
- TED 2014 EN-DE
  - Translation of TED talks
WMT – English to German

bleu

Baseline NCE SOUL SOUL+NCE

Feature

ListNet

PRO

KBMira

MERT

No Resco.
WMT – German to English

BLEU

Baseline SDWL SDWL+RBMTM

Feature

ListNet PRO KBMira MERT No Resco.
Convergence

![Graph showing convergence of BLEU+1 scores over samples (x1000) with Dev score range from 13 to 15.]
Score normalization

![Graph showing BLEU scores vs range with two curves: one for Score and one for Feature]
TED – English to German

Feature  

Baseline  extra Dev Data

BLEU

ListNet  PRO  KBMira  MERT  No Resco.
Conclusion

- Presented a new technique to train log-linear model
  - Scale to many features
  - Consider whole list
  - Technique can also be applied to more complex models
- Evaluated using different tasks and languages
  - WMT English – German
  - WMT German – English
  - IWSLT –TED English – German
- Translation quality improvements measured in BLEU score
  - Outperform MERT in all configurations
  - Less prone to overfitting
## WMT – English to German

<table>
<thead>
<tr>
<th>System</th>
<th>Baseline</th>
<th>NCE</th>
<th>SOUL</th>
<th>SOUL+NCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Test</td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>Baseline</td>
<td>20.19</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## WMT – German to English

<table>
<thead>
<tr>
<th>System</th>
<th>Baseline</th>
<th></th>
<th>SDWL</th>
<th></th>
<th>SDWL+RBMTM</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dev</td>
<td>Test</td>
<td>Dev</td>
<td>Test</td>
<td>Dev</td>
</tr>
<tr>
<td>Baseline</td>
<td>27.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MERT</td>
<td>28.18</td>
<td>27.80</td>
<td></td>
<td>28.24</td>
<td>27.65</td>
<td>28.23</td>
</tr>
<tr>
<td>KB-MIRA</td>
<td>28.23</td>
<td><strong>28.06</strong></td>
<td></td>
<td>28.18</td>
<td>28.00</td>
<td>28.00</td>
</tr>
<tr>
<td>PRO</td>
<td>27.38</td>
<td>28.01</td>
<td></td>
<td>27.56</td>
<td>28.14</td>
<td>28.68</td>
</tr>
<tr>
<td>ListNet</td>
<td>28.00</td>
<td>27.87</td>
<td></td>
<td>27.89</td>
<td><strong>28.18</strong></td>
<td>27.94</td>
</tr>
</tbody>
</table>
### TED – English to German

<table>
<thead>
<tr>
<th>System</th>
<th>Baseline Dev</th>
<th>Baseline Test</th>
<th>extra Dev Data Dev</th>
<th>extra Dev Data Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>23.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MERT</td>
<td>27.69</td>
<td>23.46</td>
<td>25.63</td>
<td>23.36</td>
</tr>
<tr>
<td>KB-MIRA</td>
<td>27.47</td>
<td>23.19</td>
<td>25.65</td>
<td>23.76</td>
</tr>
<tr>
<td>PRO</td>
<td>26.67</td>
<td>23.10</td>
<td>25.00</td>
<td>23.65</td>
</tr>
<tr>
<td>ListNet</td>
<td>27.37</td>
<td><strong>23.51</strong></td>
<td>25.49</td>
<td><strong>24.08</strong></td>
</tr>
</tbody>
</table>