Combining Coherence Models and Machine Translation Evaluation Metrics for Summarization Evaluation

Zheng Lin 1,2, Chang Liu 1, Hwee Tou Ng 1, Min-Yen Kan 1
1 National University of Singapore
2 SAP Research

Introduction

• A good machine-generated summary should have high content coverage and linguistic quality
• State-of-the-art summarization systems: Extraction-based, focusing on content
• Current AESOP task focuses on: Content, readability, and overall responsiveness
• Lin et al. (2011) used a discourse model to discern original text from its permutation
→ Adapt the model to evaluate readability
• Parallel between evaluations of MT and summarization
→ Adapt a state-of-the-art MT evaluation metric to evaluate summary content
• Combine 2 models to evaluate responsiveness with a trained regression model

TESLA: MT Evaluation Metric (Liu et al. 2010, Dahlmeier et al. 2011)

• Extends BLEU with linear programming-based matching
• Uses linguistic resources
• Considers both precision and recall
• Align 2 BNGs to maximize overall similarity

Adapting TESLA for summarization

• Mimic ROUGE-SU4: construct 1 matching problem between unigrams and 1 between skip bigrams with a window size of 4, average to give a final score
• Do not match synonyms and POS, since most systems are extraction-based
• Significance test: Koehn’s bootstrap resampling
• Tested on AESOP 2011
• Evaluated against:
  - Pearson’s r, Spearman’s p, Kendall’s t

Experiments

• Initial summarization task: outperforms all metrics on all correlations
  - Significantly better than R-2 on Pearson
• Update summarization task: ranks 2nd, 1st, and 2nd
  - Significantly better than R-2 on Pearson

Experiments

• Human judges score each model/candidate summary with a readability score from 1 to 5
  - List of training instances
• SVM® preference ranking
  - Trained on AESOP 2009 - 2010, tested on 2011
• Experiments
  - LIN: outperforms all metrics on both tasks
  - Better results on ranking-based Spearman and Kendall due to the ranking model
  - Either new feature source improves all scores
  - DICOMER: adding both gave the best performance for all scores

Koehn’s significance test

<table>
<thead>
<tr>
<th>Terms</th>
<th>P</th>
<th>S</th>
<th>K</th>
<th>P</th>
<th>S</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIN+C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DICOMER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Predicting Readability Scores

• Human judges score each model/candidate summary with a readability score from 1 to 5

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Discussion

• Initial task: correlations for content are consistently slightly higher than responsiveness
• Update task: correlations for content and responsiveness are overlapping
• Correlations for readability are much lower than those for content and readability: a gap of ~0.2
→ much room for improvement for readability
• Correlations are always better on initial task
→ eval metric needs to consider update factor

Two New Feature Sources

• Whether a relation is Explicit or Non-Explicit
  - Explicit and Non-Explicit have different distribution on each relation, e.g.
  - Exp.Arg2 to N.Exp.Arg1
• Whether one relation is embedded in another
  - Important to know how well-structured a summary is
  - Represented by multiple discourse roles in each cell

Introduction

• Lin et al. (2011)’s Coherence Model
  - Japanese normally depends heavily on the Highland Valley and Cananea mines as well as the Bougainville mine in Papua New Guinea.
  - As the Highland Valley and Cananea begin operating, they are expected to resume their roles as Japan’s suppliers.

Discussion

• DICOMER: Evaluating Summary Readability
  - A readable text should be coherent
  - An incoherent text will result in low readability
  - A coherence model can also measure readability

Lin et al. (2011)’s Coherence Model

<table>
<thead>
<tr>
<th>s_i</th>
<th>n_i</th>
<th>Comp.Arg1</th>
<th>Comp.Arg2</th>
<th>Temp.Arg1</th>
<th>Temp.Arg2</th>
<th>Exp.Arg1</th>
<th>Exp.Arg2</th>
</tr>
</thead>
<tbody>
<tr>
<td>nil</td>
<td>nil</td>
<td>nil</td>
<td>nil</td>
<td>nil</td>
<td>nil</td>
<td>nil</td>
<td>nil</td>
</tr>
<tr>
<td>Comp.Arg1</td>
<td>Comp.Arg2</td>
<td>Temp.Arg1</td>
<td>Temp.Arg2</td>
<td>Exp.Arg1</td>
<td>Exp.Arg2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discourse role transition prob of length 2 and 3:
  - e.g., Comp.Arg2→Exp.Arg2 = 2/25 = 0.08

Prediction

<table>
<thead>
<tr>
<th>Terms</th>
<th>P</th>
<th>S</th>
<th>K</th>
<th>P</th>
<th>S</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIN+C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DICOMER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CREMERS: Evaluating Overall Responsiveness

We applied SVM® to train a regression model with TESLA-S and DICOMER scores as features

• 3 kernels: linear, polynomial, radial basis
• Trained on AESOP 2009 - 2010, tested on 2011

Experiments

• RBF outperforms all AESOP metrics:
  - 1.71%, 3.86%, 4.60% on Pearson, Spearman, and Kendall
• Update task: all 3 models do not perform as well
• Koehn’s sig test: CREMERS significantly outperforms ROUGE-2 and -SU4 on initial task

Discussion

• Initial task: correlations for content are consistently slightly higher than responsiveness
• Update task: correlations for content and responsiveness are overlapping
• Correlations for readability are much lower than those for content and readability: a gap of ~0.2
→ much room for improvement for readability
• Correlations are always better on initial task
→ eval metric needs to consider update factor