Predicting Summary Quality using Limited Human Input

Annie Louis & Ani Nenkova
University of Pennsylvania
Four resource-poor methods to predict summary quality

- Evaluation using only the input
  - High input-summary similarity = better summary
  - Predicting when systems would do badly

- Evaluation using system output
  - Adding pseudo-models to human models
  - Wisdom of the crowds
    - all systems’ output make a great model
Results apply only to automatic summaries

- Numbers we report are not the officially distributed ones from the AESOP track

- Two uberbaselines—human summaries were included which invalidated the results computed

- Correlations were recomputed

- Only difference—uberbaselines excluded
TAC ’09 AESOP Data

- 44 multi-document inputs
- 2 tasks
  - Query focused
  - Update
- 53 automatic systems
  - 52 peers, 1 automatic baseline
- 2 oracle systems
  - Not used in our work
Human Scores

- **Pyramid evaluation**
  - Multiple human summaries – 4 models in TAC ’09
  - Can provide feedback about why a summary is bad
  - Significant annotation effort

- **Responsiveness scores**
  - Combined measure of content and linguistic quality
  - Direct human judgements
  - Scale 1 - 10
Comparing predictions with human judgements

- **System-level ~ which system is better overall?**
  - Average predicted scores for a system over the test set
  - Average human scores
  - Correlation between rankings

- **Input-level ~ which summary is better for an input?**
  - Correlation between rankings of summaries for each individual input
  - % of inputs with significant correlations
1. Input-summary similarity

- Evaluate content selection using **no human models** at all
Intuitive measure of summary quality

- Evaluation on non-standard test sets
  - With no model summaries

- Likely to be a good objective function for content selection

- But many ways to measure similarity
  - KL, JS divergence
  - Cosine similarity
  - Topic word similarity
  - Frequency based summary likelihood
UPenn at TAC ‘08

- Analysis of different input-summary similarity metrics [TAC ‘08, EMNLP ‘09]

- Performance varies with different features
  - Best features ~ information-theoretic measures
  - Worst ~ frequency based metrics

- Top features were highly predictive of human scores
  - Best correlation at system-level ~ 0.89
Best predictor - Jensen Shannon divergence

- Distance between 2 probability distributions
  - As average KL divergence from their mean distribution

Low divergence ~ better summary quality

\[
JS (Inp \parallel Summ) = \frac{1}{2} \left[ KL (Inp \parallel A) + KL (Summ \parallel A) \right]
\]

\[
A = \frac{Inp + Summ}{2}, \text{ mean distribution of Input and Summary}
\]
Regression metric

- A range of distributional similarity and other features
  - KL divergence
  - JS divergence
  - Cosine similarity
  - Topic signature based features
  - Summary likelihood under a frequency based model
Top 2 features on ’09 – validated findings from last year

<table>
<thead>
<tr>
<th></th>
<th>Query Task</th>
<th>Update Task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pyramid</td>
<td>Resp.</td>
</tr>
<tr>
<td>JS divergence</td>
<td>-0.74</td>
<td>-0.71</td>
</tr>
<tr>
<td>Regression</td>
<td>0.77</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Best performance on ’08: JS = 0.89 with pyramid scores

- Good content can be predicted from the input
- Information-theoretic features provide good estimates
2. Summarization difficulty of the source text

- Average system performance on an input can be predicted with good accuracies
Some inputs are more difficult for systems

- Systems ignore properties of individual inputs
  - Very low average performance on certain inputs

- Input difficulty can be measured by a number of features [ACL ’08, EMNLP ’09]

- Can predict when average system performance will be below the mean value
Defining what is easy/difficult for systems

- Difficult input
  - Most systems perform poorly
  - Low average system score

- 2 classes – easy, difficult
  - Above/below mean average system score
  - Equal number of inputs in both classes
Good indicators of difficult inputs

- Large vocabulary size
- Fewer descriptive words – hard to identify through frequency and repetition
- Low redundancy between input documents
- No clear topic

- 6 significant features
- Good accuracies in identifying difficult inputs
  - 10% above baseline
### Predictions on TAC ‘09 data

<table>
<thead>
<tr>
<th></th>
<th>All inputs</th>
<th>Extremes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query task</td>
<td>61.36</td>
<td>60.00</td>
</tr>
<tr>
<td>Update task</td>
<td>59.09</td>
<td>75.00</td>
</tr>
</tbody>
</table>

* Extremes – 10 each most easy and difficult

Trained on DUC 2002-2004

- Properties of input predictive of average system performance
- Specialized content selection necessary to smooth out variations
3. System summaries + Human models

- Pseudo-models for summary evaluation
System level – one model is enough

- Another likely setup on non-standard test sets
- Robust system-level rankings on large test sets

<table>
<thead>
<tr>
<th></th>
<th>Query Task</th>
<th>Update Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSU4-recall</td>
<td>Pyramid</td>
<td>Resp.</td>
</tr>
<tr>
<td>1 model</td>
<td>0.92</td>
<td>0.80</td>
</tr>
<tr>
<td>4 models</td>
<td>0.92</td>
<td>0.79</td>
</tr>
<tr>
<td>Pyramid</td>
<td>Resp.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.80</td>
<td>0.69</td>
</tr>
<tr>
<td>0.85</td>
<td>0.69</td>
<td></td>
</tr>
</tbody>
</table>
Choose one model per input
- Alphabetical order of model name

Considerably fewer inputs with significant correlations

<table>
<thead>
<tr>
<th></th>
<th>Query Task</th>
<th>Update Task</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RSU4-recall</strong></td>
<td>Pyramid</td>
<td>Resp.</td>
</tr>
<tr>
<td>1 model</td>
<td>84.09</td>
<td>79.54</td>
</tr>
<tr>
<td>4 models</td>
<td>95.45</td>
<td>81.82</td>
</tr>
</tbody>
</table>
Can we improve the evaluation using system output? [Albrecht & Hwa ‘08]

- Related work in Machine translation
- One human reference translation
- Off-the-shelf systems as pseudo-references
- Features to compare other translations with pseudo-references
- Regression based scoring
- Improved correlations compared to using a single human reference
“Pseudo-model” system summaries

- Pseudo-model ~ systems predicted to be best using available model summary

- Compute ranks based on the human model
- Treat top systems as “pseudo-models”
Two selection methods

- **Global**
  - System level ranking using RSU4
  - Select top 3 systems as pseudo-models

- **Local**
  - Use top 3 systems for each input as pseudo-models

- **Final rankings**
  - JS divergence with 1 model + 3 pseudo-models
## Mixed results

<table>
<thead>
<tr>
<th></th>
<th>Query Task</th>
<th>Update Task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pyramid</td>
<td>Resp.</td>
</tr>
<tr>
<td>1 human</td>
<td>84.09</td>
<td>79.54</td>
</tr>
<tr>
<td>Global sel.</td>
<td>93.18</td>
<td>79.55</td>
</tr>
<tr>
<td>Local sel.</td>
<td>93.18</td>
<td>75.00</td>
</tr>
</tbody>
</table>

- Improvements for pyramid
- Not much gains for responsiveness
- On ’08 data, local selection was better
4. System summaries only

- Collection of system summaries is useful for evaluation
Can system summaries alone be used for evaluation?

- Similar to the pyramid method
  - Common content across multiple human summaries more important

- Different systems ~ different content selection methods
  - Agreement among systems ~ very important content

- Collection of system summaries as a model
  - Indicative of what is important?
System summary based evaluation

- Divergence from vocabulary distribution of system summaries

Collective vocabulary of all system summaries

Vocabulary distribution of individual system summary

Low divergence ~ higher scores
## Very high correlations with human scores

<table>
<thead>
<tr>
<th>System-level</th>
<th>Query Task</th>
<th>Update Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>System summaries</td>
<td>Pyramid</td>
<td>Resp.</td>
</tr>
<tr>
<td></td>
<td>-0.93</td>
<td>-0.81</td>
</tr>
<tr>
<td>RSU4 – 4 models</td>
<td>0.92</td>
<td>0.79</td>
</tr>
</tbody>
</table>

- Percentage of inputs with significant correlations
  - 77 to 90%

- Collective knowledge of systems is useful
  - Possibility of system combination for summarization
Conclusions

- 4 methods to predict summary quality that use very little or no human input

- Based upon system summaries
  - Pseudo-models: help only for pyramid correlations
  - Collection of system summaries: very indicative of good content

- Based upon the input
  - Input-summary similarity: highly predictive
  - Input difficulty features: predictive of average system performance
References

- Automatically Evaluating Content Selection in Summarization without Human Models
  - Annie Louis & Ani Nenkova, EMNLP 2009

- Performance Confidence Estimation for Automatic Summarization
  - Annie Louis & Ani Nenkova, ACL 2009

- Summary Evaluation without Human Models
  - Annie Louis & Ani Nenkova, TAC 2008

- Can you summarize this? Identifying correlates of input difficulty for generic multi-document summarization
  - Ani Nenkova & Annie Louis, ACL-HLT 2008