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

Penn
UNIVERSITY of PENNSYLVANIA

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

Penn
UNIVERSITY of PENNSYLVANIA

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

Penn
UNIVERSITY of PENNSYLVANIA

Best predictor - Jensen Shannon divergence

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

Low divergence ~ better summary quality

$$JS(Inp \parallel Summ) = \tfrac{1}{2}[\,KL(Inp \parallel A) \; + \; KL(Summ \parallel A)]$$

$$A = \frac{Inp + Summ}{2}, \; mean \;\; distributi \;\; on \;\; of \;\; Input \;\; and \;\; Summary$$

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

Penn
University of Pennsylvania

Top 2 features on '09 – validated findings from last year

	Query Task		Update Task	
	Pyramid	Resp.	Pyramid	Resp.
JS divergence	-0.74	-0.71	-0.72	-0.61
Regression	0.77	0.67	0.71	0.54

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

Penn
University of Pennsylvania

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

	All inputs	Extremes
Query task	61.36	60.00
Update task	59.09	75.00

* 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

	Query Task		Update Task	
RSU4-recall	Pyramid	Resp.	Pyramid	Resp.
1 model	0.92	0.80	0.80	0.69
4 models	0.92	0.79	0.85	0.69

Input-level – more models necessary

- Choose one model per input
  - Alphabetical order of model name

- Considerably fewer inputs with significant correlations

Query Task

RSU4-recall	Pyramid	Resp.
1 model	84.09	79.54
4 models	95.45	81.82

Update Task

Pyramid	Resp.
86.36	75.00
100	86.36

Penn
UNIVERSITY of PENNSYLVANIA

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

	Query Task: Pyramid	Query Task: Resp.	Update Task: Pyramid	Update Task: Resp.
1 human	84.09	79.54	86.36	75.00
+ 3 pseudo: Global sel.	93.18	79.55	93.18	77.27
+ 3 pseudo: Local sel.	93.18	75.00	86.36	79.55

- 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?

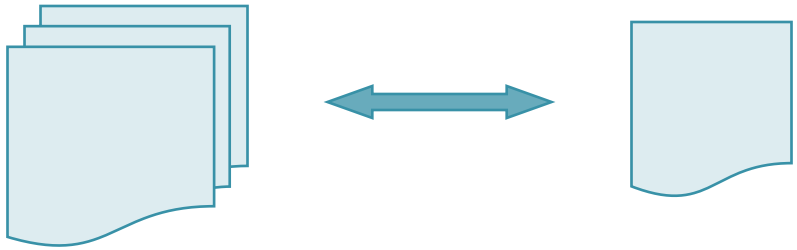
- Similar to the pyramid method
  - Common content across multiple human summaries more important
- Different systems ~ different content selection methods
  - Agreement among systems ~ very important content
- Collection of system summaries as a model
  - Indicative of what is important?

Penn
UNIVERSITY of PENNSYLVANIA

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

	Query Task		Update Task	
System-level	Pyramid	Resp.	Pyramid	Resp.
System summaries	-0.93	-0.81	-0.89	-0.79
RSU4 – 4 models	0.92	0.79	0.85	0.69

- 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

Penn
UNIVERSITY of PENNSYLVANIA

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

Penn
University of Pennsylvania