Automatic Evaluation of Linguistic Quality in Multi-Document Summarization

Pitler, Louis, Nenkova 2010
Presented by Dan Feblowitz and Jeremy B. Merrill
Motivation

- Automatic evaluation of content selection is already done.
  - ROUGE: automated metric for info content. (Lin and Hovy, 2003; Lin, 2004)

- No automatic evaluation of linguistic quality available.
  - We want fluent and easy-to-read summaries.
  - How to test?
Intuitions: Aspects of Linguistic Quality

- **Grammaticality**
  - *The Police found no second armed man. LOS ANGELES -- A sniping incident Sunday damaged helicopter.*

- **Non-redundancy**
  - *Bill Clinton ate a banana yesterday, Bill Clinton liked it. Bill Clinton was in Los Angeles.*

- **Referential Clarity**
  - *The beer savvy participant, a 20-year-old male, was arrested Saturday. “This was really irresponsible,” she said.*

- **Focus**
  - *To show solidarity with dining hall workers, Bill Clinton ate a banana. He was at Frary. Frary contains a mural by some Mexican muralist.*

- **Structure and Coherence**
  - *Harvey Mudd was founded in 1954. It is a engineering college. It has eight dorms. Its founder was named Harvey.*
Correlation Among Aspects

- Referential Clarity, Focus and Structure are significantly correlated with each other. (Along with a few more significant correlations.)

- Linguistic quality rankings correlate positively with content quality rankings.

- Human rankers.
Goal

- Find automated measures that correlate with the intuition-based aspects.
  - System-level evaluation
  - Input-level evaluation
Automated Measures

- Language Modeling: Gigaword corpus /1-,2-,3-gram
- Entity explanation: Named Entities, NP Syntax
- Cohesive devices: demonstratives, pronouns, definite descriptions, sentence-initial discourse connectives
- Sentence fluency: length, fragments, etc.
- Coh-Metrix: Psycholinguistic readability measures
- Word Coherence
  - Treat adjacent sentences as parallel texts
  - Calculate “translation model” in each direction
Automated Measures (cont)

• Continuity
  • **Summarization specific**: Measures likelihood that discourse connectives retain their context. Does previous sentence in summary match previous sentence in input?
  • **Cosine similarity** of words across adjacent sentences.
  • **Coreference**: Pronoun resolution system. Probability of antecedent presence in sentence, previous sentence.

• Entity coherence
  • Matrix of entities’ grammatical roles; measure transition probabilities among entity’s role in adjacent sentence.
Experiment Setup

- Data from summarization task of 2006/2007 Document Understanding Conference
  - 2006 (training/dev sets) 50 inputs, 35 systems tested
    - Jackknifing
  - 2007 (test set) 45 inputs, 32 systems

- One ranker for each feature group, plus meta-ranker.

- Rank systems/summaries relative to a gold standard human ranking based on each automated measure.

- Find correlations with human ranking on aspects.
Results (System-Level)

- Prediction Accuracy
  - Percentage of pairwise comparisons matching gold standard.
  - Baseline: 50% (random)

- System-level: (for summarization system)
  - Prediction accuracies around 90% for all aspects
  - Sentence fluency method single best correlation with Grammaticality. Meta-ranker has best overall correlation.
  - Continuity method best correlates with Non-Redundancy, Referential Clarity, Focus, Structure.
Results (Input-Level)

- Input-level: (for each summary)
  - Prediction accuracies around 70% -- harder task.
  - Sentence fluency method single best correlation with grammaticality.
  - Coh-Metrix single best correlation with Non-Redundancy
  - Continuity best correlates with Referential Clarity, Focus, Structure.
  - Meta-ranker best correlation for all aspects.
Results (Human-Written)

- Input-level analysis on human-written, abstractive summaries.
  - Abstractive: Rewritten content
  - Extractive: Extracts subset of content, i.e. picking sentences

- Grammaticality: NP Syntax (64.6%)
- Non-redundancy: Coherence devices (68.6%)
- Referential Clarity: Sentence Fluency, Meta-Ranker (80.4%)
- Focus: Sentence Fluency, LMs (71.9%)
- Structure: LMs (78.4%)
Components of Continuity

Subsets of features in continuity block removed one-at-a-time to measure effect of each.

Cosine similarity had greatest effect (-10%)

Summary-specific features were second (-7%)

Removing coreference features had no effect.
Conclusions

• Continuity features correlate with linguistic quality of machine-written summaries.

• Sentence fluency features correlate with grammaticality.

• LM and entity coherence features also correlate relatively strongly.

• This will make testing systems easier. Hooray!
Questions?