Content Realization: Linguistic Quality

Ling 573
Systems and Applications
May 9, 2017

Roadmap

- Content realization in summarization
  - Goals
  - Broad approaches
- Readability and linguistic quality:
  - Corpus study and analysis
  - Automatic evaluation
  - Improvements for MDS

Goals of Content Realization

- Abstractive summaries:
  - Content selection works over concepts
  - Need to produce important concepts in fluent NL
- Extractive summaries:
  - Already working with NL sentences
  - Extreme compression: e.g 60 byte summaries: headlines
  - Increase information:
    - Remove verbose, unnecessary content
    - More space left for new information
  - Increase readability, fluency, linguistic quality
    - Present content from multiple docs, non-adjacent sents
  - Improve content scoring
    - Remove distractors, boost scores: i.e. % signature terms in doc

Broad Approaches

- Abstractive summaries:
  - Complex Q-A: template-based methods
  - More generally: full NLG: concept-to-text
- Extractive summaries:
  - Sentence compression:
    - Remove “unnecessary” phrases:
      - Information? Readability?
  - Sentence reformulation:
    - Reference handling
      - Information? Readability?
  - Sentence fusion: Merge content from multiple sents

Linguistic Quality

Evaluation

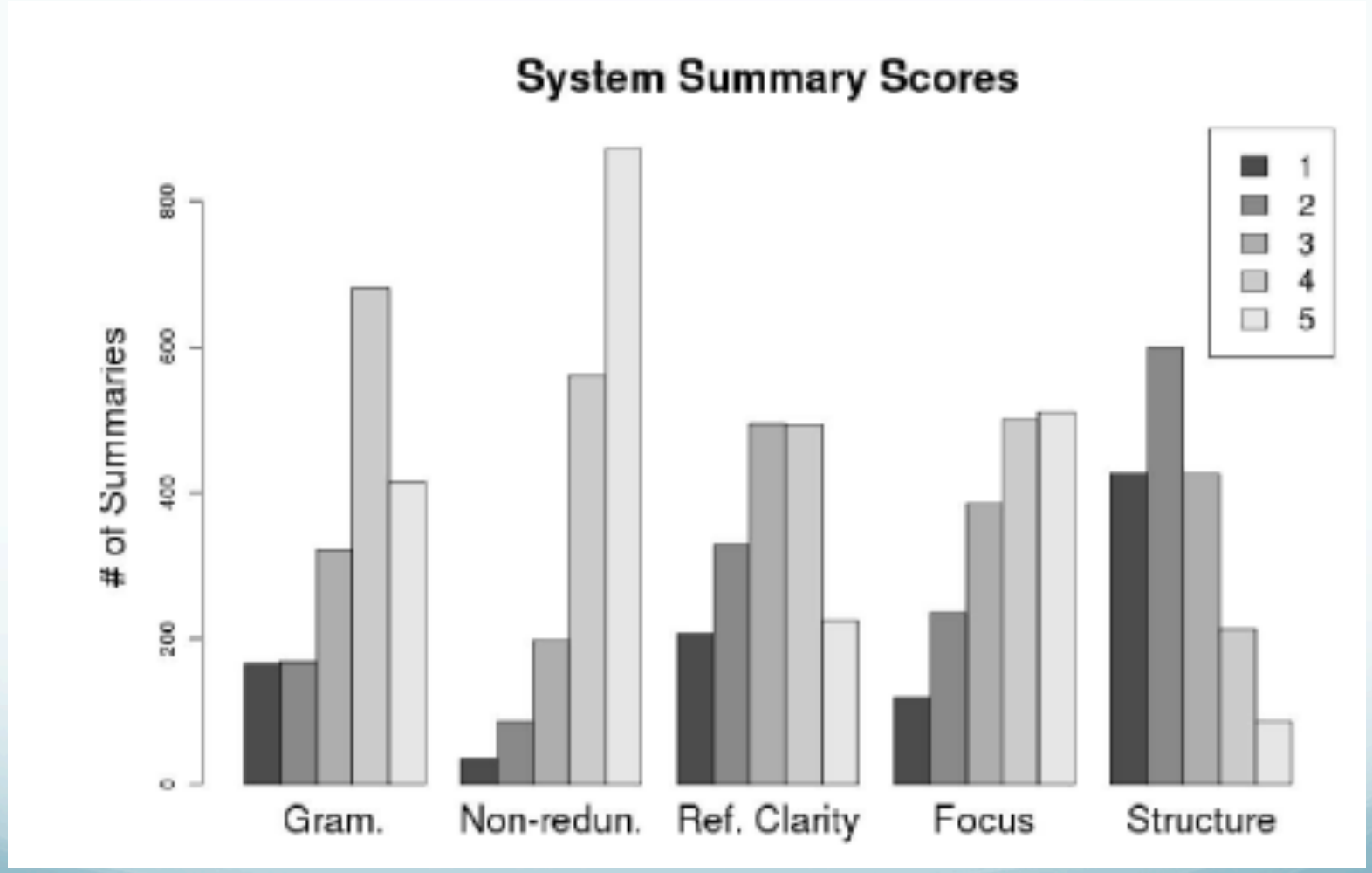
- Shared tasks:
  - Take content as primary evaluation measure
    - ROUGE, Pyramid, (manual) Responsiveness
  - Linguistic quality also part of formal evaluation

- TAC "Readability":
  - Scored manually on 5-point Likert scale
  - Aims to capture readability, fluency
    - Independent of summary content

What is “Readability”?

- According to TAC,
- Assessors consider (and rate 1-5) each of:
  - Grammaticality:
    - No fragments, datelines, ill-formed sentences, etc
  - Non-redundancy:
    - No unnecessary repetition: includes content, sentences, or full NPs when pronoun is better
  - Referential clarity:
    - Both presence/salience of antecedents, relevance of items
  - Focus:
    - Only content related to summary
  - Coherence: “Well-structured”

Score Distributions



What is "Readability"? II

- Definition subsumes many phenomena, errors
- What types of errors do these systems make?
- What errors, issues are reflected in the scores?
- LQVSumm (Friedrich et al, 2013)
  - Annotate linguistic "violations" in automatic summaries
    - TAC2011 data: ~2000 "peer" summaries
    - Categorize and tabulate
  - Assess correlation with Readability scores

Example

Charles Carl Roberts IV may have planned to molest the girls at the Amish school, but police have no evidence that he actually did. Charles Carl Roberts IV entered the West Nickel Mines Amish School in Lancaster County and shot 10 girls, killing five. The suspect apparently called his wife from a cell phone shortly before the shooting began, saying ***he was "acting out in revenge for something that happened 20 years ago,*** *Miller said. The gunman, a local truck driver Charles Roberts,* ***was apparently acting in "revenge" for an incident that happened to him 20 years ago.***

Friedrich et al, 2013, p. 1591, Fig. 1

Violation Categories

- Entity mentions:
  - Affect coreference and readability
  - FM_EXPL: First mention w/o explanation
  - SM+EXPL: Subsequent Mention w/explanation
  - DNP_REF: Definite NP w/o previous mention
  - INP+REF: Indefinite NP w/ previous mention
  - PRN+MISSA: Pronoun w/missing antecedent
  - PRN+MISSLA: Pronoun w/misleading antecedent
  - ACR_EXPL: Acronym w/o explanation

Violation Categories

- Clausal level:
  - Arbitrary spans – up to sentence level
  - INCOMPLSN: Incomplete sentence
  - INCLDATE: included dateline info
  - OTHRUNGR: other ungrammatical
  - NOSEMREL: No semantic relation b/t sentences
  - NODISREL: Discourse relation doesn't fit
  - REDUNINF: Redundant information

violation type	count	avg/doc	Pearson's *r*		
			Readability	Pyramid	Respons.
entity level violations					
DNP-REF	958	0.50	**-0.122**	**-0.166**	**-0.133**
FM-EXPL	792	0.41	0.006	-0.050	**-0.066**
INP+REF	430	0.22	-0.052	**0.235**	**0.109**
PRN+MISSA	361	0.19	**-0.191**	**-0.140**	**-0.156**
SM+EXPL	162	0.08	0.020	**0.089**	0.045
PRN+MISLA	27	0.01	**-0.065**	**-0.073**	**-0.089**
ACR-EXPL	11	0.01	-0.038	-0.056	-0.006
sum(DNP-REF, PRN+MISSA)	1319	0.68	**-0.204**	**-0.208**	**-0.192**
sum(entity level violations)	2741	1.42	**-0.167**	-0.074	**-0.127**
clause level violations					
INCOMPLSN	1,044	0.54	**-0.210**	0.000	-0.029
OTHRUNGR	793	0.41	**-0.180**	0.007	-0.016
INCLDATE	412	0.21	**-0.090**	0.039	0.051
REDUNDINF	504	0.26	**-0.160**	**0.156**	**0.077**
NOSEMREL	142	0.07	**-0.148**	**-0.102**	**-0.132**
NODISREL	91	0.05	-0.025	**-0.081**	**-0.062**
misleading discourse connectives⋆	114	0.06	-	-	-
sum(clause level violations)	2,986	1.54	**-0.325**	0.041	-0.016
sum(clause level violations, DNP-REF, PRN+MISSA)	4,305	2.22	**-0.385**	**-0.084**	**-0.122**
sum(all violations)	5,727	2.96	**-0.356**	-0.022	**-0.101**

Friedrich et al. 2013, p. 1596, Tab. 2

Further Analysis

- Linear model investigates the relationship of particular errors to readability

Feature	Weight	Feature	Weight
Intercept	3.407	DNP-REF	-0.157
ACR-EXPL	-0.361	OTHRUNGR	-0.155
PRN+MISLA	-0.355	INCLDATE	-0.151
INCOMPLSN	-0.275	INP+REF	-0.067
NOSEMREL	-0.262	NODISREL	-0.046
REDUNDINF	-0.259	FM-EXPL	-0.023
PRN+MISSA	-0.236	SM+EXPL	0.038

- Most significant factors: Missing/Misleading refs, fragments, redundant content, poor coherence
- Total # of errors well-correlated with system ranks

Friedrich et al, 2013, p. 1596, Tab. 3

Automatic Evaluation of Linguistic Quality

- Motivation:
  - No focus on linguistic quality b/c no way to tune to it
  - Everyone uses ROUGE b/c you can tune
    - Explicitly tuned in many ML models

- Alternative strategies:
  - Micro: Learn to predict component scores
  - Macro: Learn to predict overall readability score
    - Intuitively: error count (LQVSumm) predicts well, but…
      - Errors manually derived

Micro-Quality Prediction

- (Pitler et al, 2010) via SVM ranking

- Evaluate multiple measures aimed to model LQ
  - General word choice, sequence: Language Models
  - Reference form:
    - Named Entities:
      - Modifiers for 1st mention of PERSON
      - Proportion of summary NER first mentions originally non-first
    - NP syntax: POS, phrase tags in NPs
  - Local coherence devices:
    - Count of demonstratives, pronouns, definite descriptions, and sentence initial discourse connectives

Micro-Quality Prediction

- Evaluate multiple measures aimed to model LQ
  - Continuity:
    - For each cohesive device, are sentences adjacent in source?
    - Position and confidence of antecedents of pronouns
    - Max, min, and average cosine similarity b/t sentences
  - Sentence fluency:
    - Shallow syntax features correlated w/MT quality
  - Coh-Metrix:
    - Set of psycholinguistically-based coherence feats, LSA sim
  - Word coherence: cross-sentence word cooccurrence patterns
  - Entity coherence: via Entity-grids (Brown toolkit)

Results

- System level

Feature set	Gram.	Redun.	Ref.	Focus	Struct.
Lang. models	87.6	83.0	91.2	85.2	86.3
Named ent.	78.5	83.6	82.1	74.0	69.6
NP syntax	85.0	83.8	87.0	76.6	79.2
Coh. devices	82.1	79.5	82.7	82.3	83.7
Continuity	88.8	**88.5**	**92.9**	**89.2**	**91.4**
Sent. fluency	**91.7**	78.9	87.6	82.3	84.9
Coh-Metrix	87.2	86.0	88.6	83.9	86.3
Word coh.	81.7	76.0	87.8	81.7	79.0
Entity coh.	90.2	88.1	89.6	85.0	87.1
Meta ranker	**92.9**	87.9	91.9	87.8	90.0

- Summary level

Feature set	Gram.	Redun.	Ref.	Focus	Struct.
Lang. models	66.3	57.6	62.2	60.5	62.5
Named ent.	52.9	54.4	60.0	54.1	52.5
NP Syntax	59.0	50.8	59.1	54.5	55.1
Coh. devices	56.8	54.4	55.2	52.7	53.6
Continuity	61.7	62.5	**69.7**	**65.4**	**70.4**
Sent. fluency	**69.4**	52.5	64.4	61.9	62.6
Coh-Metrix	65.5	**67.6**	67.9	63.0	62.4
Word coh.	54.7	55.5	53.3	53.2	53.7
Entity coh.	61.3	62.0	64.3	64.2	63.6
Meta ranker	**71.0**	**68.6**	**73.1**	**67.4**	**70.7**

Pitler et al, 2010, p. 550, Tab 2; p. 551, Tab 3

Findings

- Overall accuracies quite good
- Systems overall easier to rank than particular input
  - Smoothes variance, larger sample
- Continuity related features best across components
  - Ensemble of ordering, coref, cosine similarity cues
    - Though LSA-based system detects redundancy well
- Specifically tuned fluency scorer works on fluency

Macro-Quality Prediction

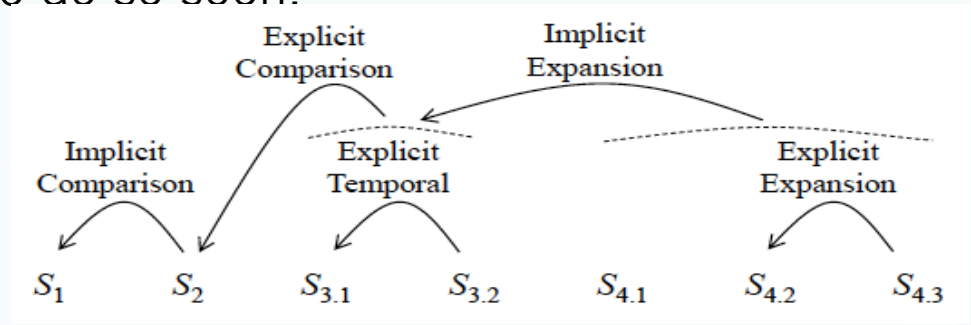
- (Lin et al, 2012) Downloadable

- High-level idea:
  - Discourse version of entity grid
    - Columns: entities (same head)
    - Rows: sentences
    - Cell values: PDTB Discourse Relation.Arg# tuples

- Variants:
  - Inter-cell sequence frequencies
    - + Additional tuples: {Non--}Explicit.Relation.Arg#
    - + Intra-cell “sequences”

(Lin et al, 2012; p. 1010; Fig 1,2; Tab 2

S_1: Japan normally depends heavily on the Highland Valley and Cananea mines as well as the Bougainville mine in Papua New Guinea.
S_2: Recently Japan has been buying copper elsewhere.
$S_{3.1}$:But as Highland Valley and Cananea begin operating,
$S_{3.2}$: they are expected to resume their roles as Japan's suppliers.
$S_{4.1}$: According to Fred Demler, metals economist for DBL, New York,
$S_{4.2}$: "Highland Valley has already started operating
$S_{4.3}$: and Cananea is expected to do so soon."



S#	Copper	Cananea	operat	depend	..
S_1	Nil	Comp.A1	Nil	Comp.A1	
S_2	Comp.A2 Comp.A1	Nil	Nil	Nil	
S_3	Nil	Comp.A2 Temp.A1 Exp.A1	Comp.A2 Temp.A1 Exp.A1	nil	
S_4	Nil	Exp.A1	Exp.A1 Exp.A2	nil	

Results

- Very strong correlations w/manual readability score
  - Beats prior predictors

Measure	Pearson	Spearman
Rouge-2	0.7524	0.3975
TAC system 6	0.8194	0.4937
DiscRelGrid	0.8556	0.6593
DiscRelGrid + Explicit tags + Within cell transitions	0.8666	0.7122