Project 1 Deliverable 5

LING 575: Summarization
Team 2

Team 2

Ben Cote

Mohamed Elkamhawy

Karl Haraldsson

Alyssa Vecht

Josh Warzecha

Project 1 Overview

Multi-document summarization

- News articles
- Various categories and topics

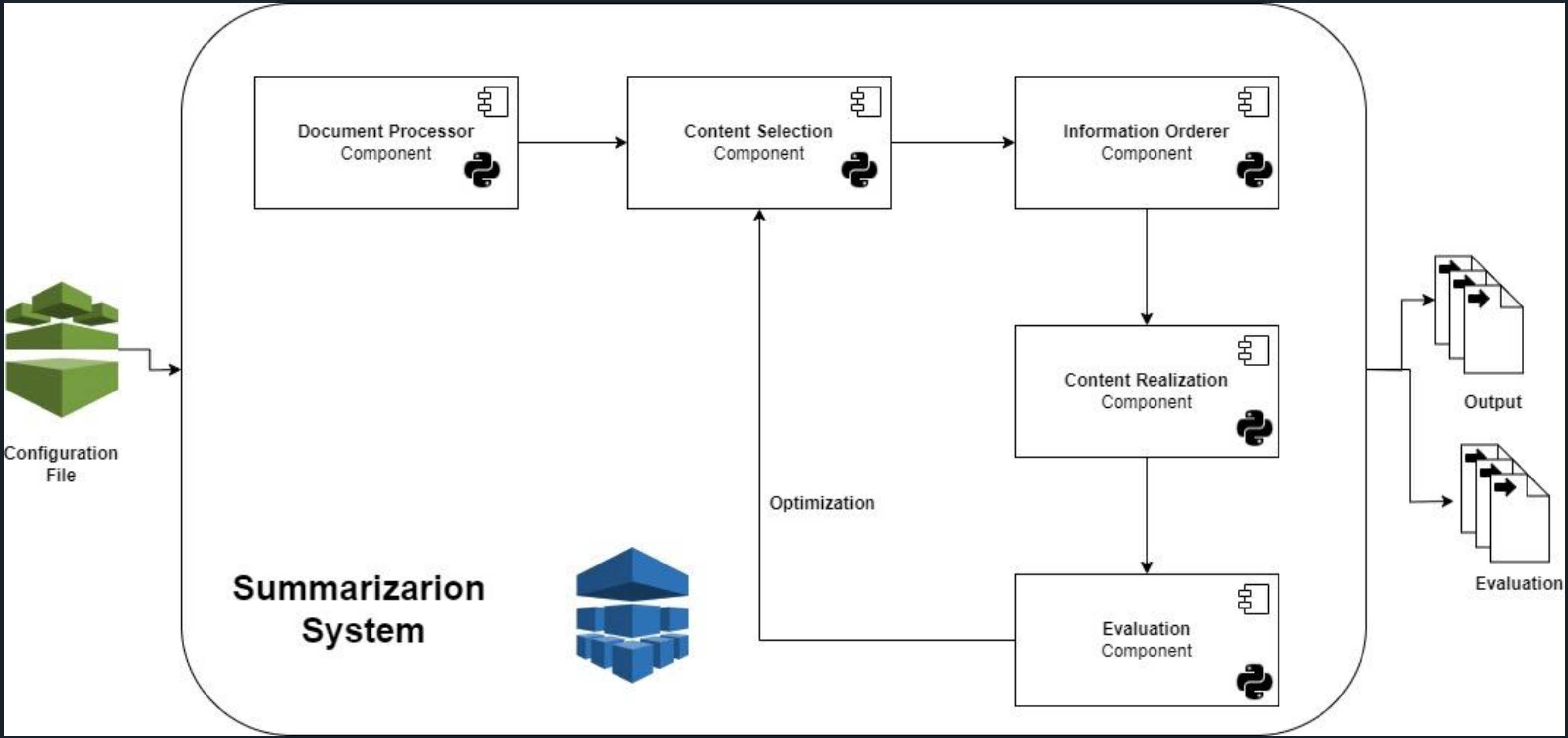
End-to-end system

- Document ingestion
- Content selection
- Information ordering
- Content realization

Evaluation of output summaries

D5: Final System

System Overview



Content Selection

Goal:

Refine the initial content selection components implemented in D4, including a mechanism for selecting content for topic-focused summarization.

Implementation:

TextRank

Use BERT embeddings as the vector representations of selected sentences (only sentences of 8 words or more). Then calculate cosine similarity between sentence representations and use pagerank algorithm to select top sentences

Topic-Focused Summarization

Incorporating Sentence-BERT for semantic embeddings and adapting LexRank to emphasize topic relevance, significantly aligning summaries with specified topics through calculated semantic similarities.

Information Ordering

Goal:

Refine the initial information ordering components implemented in D4 to improve the coherence and readability of summaries by determining an optimal sequence for the selected content.

Implementation:

CLASSY (Conroy et al, 2006):

Use MASI distance to measure similarity between sentences and apply a two-opt algorithm for optimization.

Entity-Grid (Barzilay and Lapata, 2008):

Build a logistic regression model from the training data to model the transition of entities between sentences and use this to predict the most likely permutation of sentences, incorporating dependency parsing.

Content Realization

Goal:

Compose selected and ordered content into a coherent summary that retains the integrity of the underlying sentences.

Implementation:

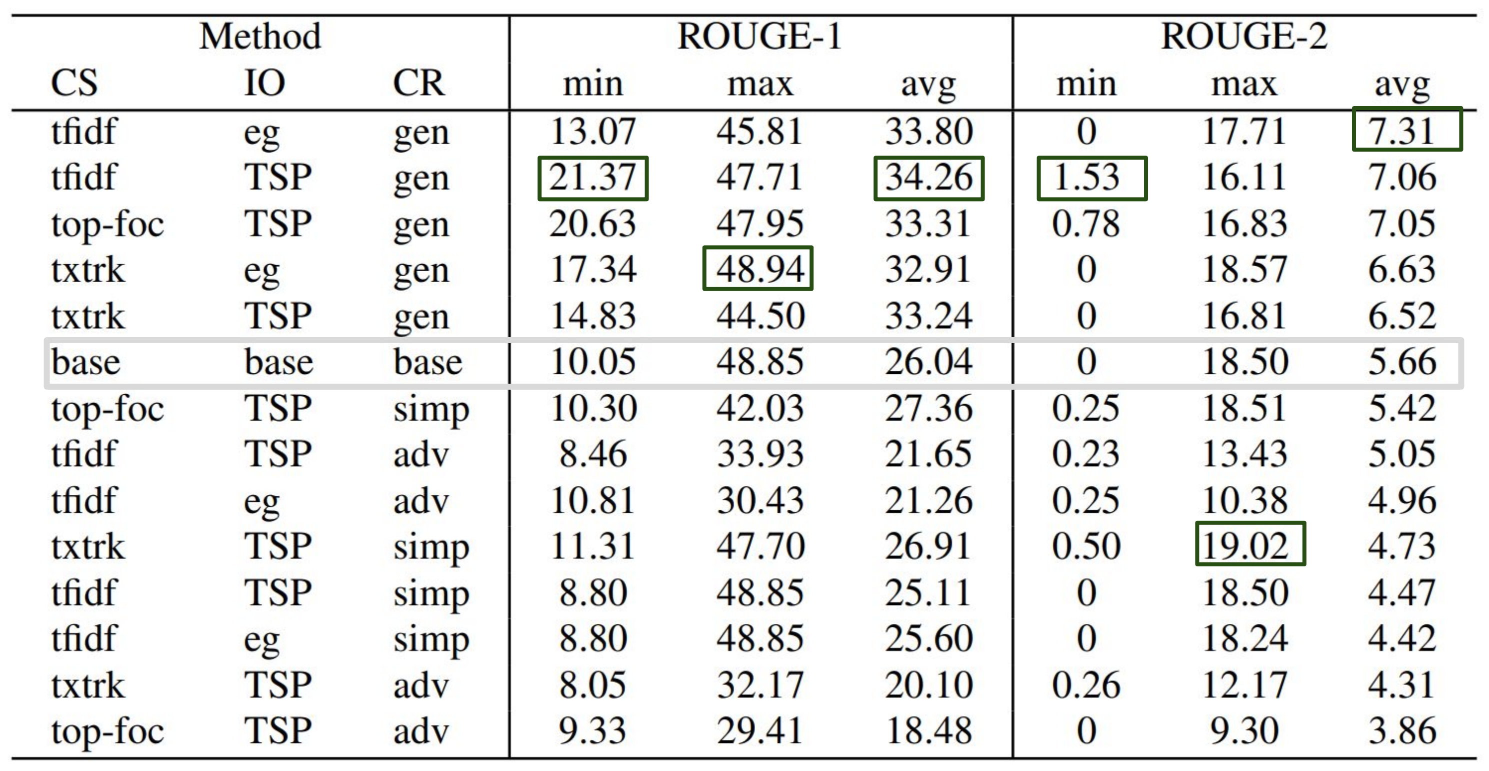
Seq2Seq Compression (Nayeem et al, 2019):

Use a seq2seq model to compress input sentences such that they fit under the 100 word limit. Use hugging face transformers summarization pipeline object with t5-base.

Generative Realization

Use generative inference components to revise selected and ordered content (OpenAI API with GPT-3.5-turbo). Instruct via prompting to ensure readability, coreference resolution, and brevity.

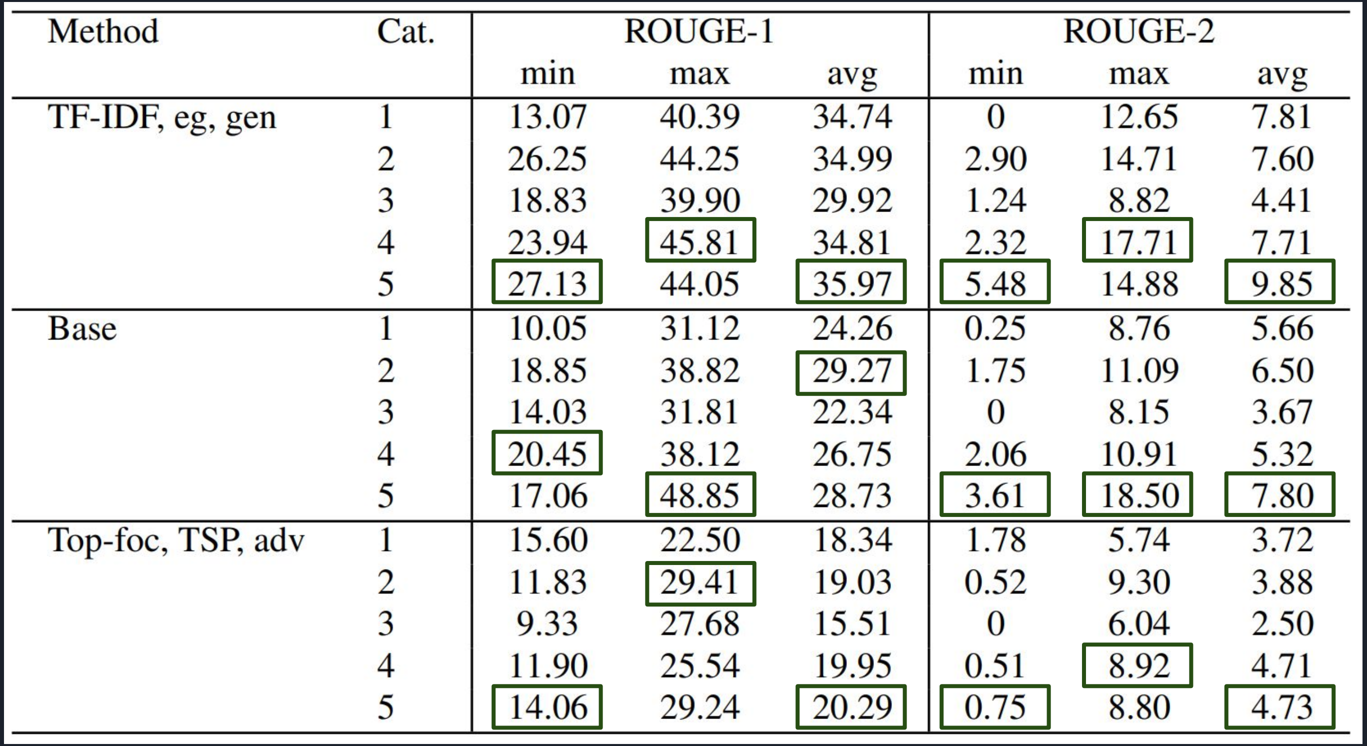
Issues & Successes



Method			ROUGE-1			ROUGE-2		
CS	IO	CR	min	max	avg	min	max	avg
tfidf	eg	gen	13.07	45.81	33.80	0	17.71	7.31
tfidf	TSP	gen	21.37	47.71	34.26	1.53	16.11	7.06
top-foc	TSP	gen	20.63	47.95	33.31	0.78	16.83	7.05
txtrk	eg	gen	17.34	48.94	32.91	0	18.57	6.63
txtrk	TSP	gen	14.83	44.50	33.24	0	16.81	6.52
base	base	base	10.05	48.85	26.04	0	18.50	5.66
top-foc	TSP	simp	10.30	42.03	27.36	0.25	18.51	5.42
tfidf	TSP	adv	8.46	33.93	21.65	0.23	13.43	5.05
tfidf	eg	adv	10.81	30.43	21.26	0.25	10.38	4.96
txtrk	TSP	simp	11.31	47.70	26.91	0.50	19.02	4.73
tfidf	TSP	simp	8.80	48.85	25.11	0	18.50	4.47
tfidf	eg	simp	8.80	48.85	25.60	0	18.24	4.42
txtrk	TSP	adv	8.05	32.17	20.10	0.26	12.17	4.31
top-foc	TSP	adv	9.33	29.41	18.48	0	9.30	3.86

System performance on devtest with varying methods for content selection (CS), information ordering (IO), and content realization (CR).

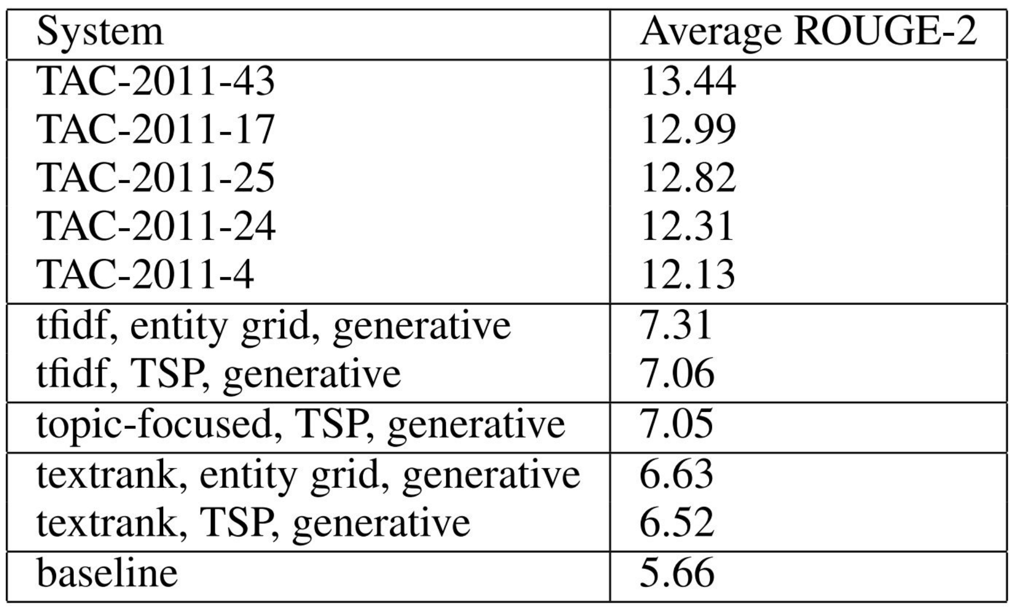
Issues & Successes



Method	Cat.	ROUGE-1			ROUGE-2		
		min	max	avg	min	max	avg
TF-IDF, eg, gen	1	13.07	40.39	34.74	0	12.65	7.81
	2	26.25	44.25	34.99	2.90	14.71	7.60
	3	18.83	39.90	29.92	1.24	8.82	4.41
	4	23.94	45.81	34.81	2.32	17.71	7.71
	5	27.13	44.05	35.97	5.48	14.88	9.85
Base	1	10.05	31.12	24.26	0.25	8.76	5.66
	2	18.85	38.82	29.27	1.75	11.09	6.50
	3	14.03	31.81	22.34	0	8.15	3.67
	4	20.45	38.12	26.75	2.06	10.91	5.32
	5	17.06	48.85	28.73	3.61	18.50	7.80
Top-foc, TSP, adv	1	15.60	22.50	18.34	1.78	5.74	3.72
	2	11.83	29.41	19.03	0.52	9.30	3.88
	3	9.33	27.68	15.51	0	6.04	2.50
	4	11.90	25.54	19.95	0.51	8.92	4.71
	5	14.06	29.24	20.29	0.75	8.80	4.73

Performance of best, worst, and baseline systems on devtest by text category.
(Top- and bottom-performing systems chosen based on ROUGE-2.)

Final Evaluation (evaltest)



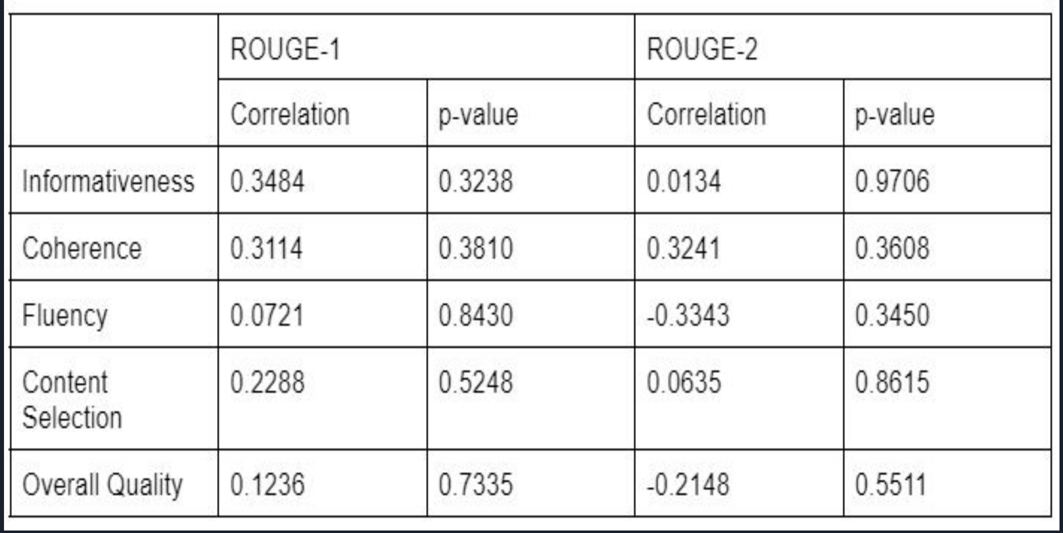
System	Average ROUGE-2
TAC-2011-43	13.44
TAC-2011-17	12.99
TAC-2011-25	12.82
TAC-2011-24	12.31
TAC-2011-4	12.13
tfidf, entity grid, generative	7.31
tfidf, TSP, generative	7.06
topic-focused, TSP, generative	7.05
textrank, entity grid, generative	6.63
textrank, TSP, generative	6.52
baseline	5.66

Comparison of our best systems, TAC-2011 shared task best systems, and baseline on evaltest.
(Top- and bottom-performing systems chosen based on ROUGE-2.)

Human Evaluation

Human Evaluation metrics:

- **Informativeness:**
  - Does the summary capture the key information of the issue?
- **Coherence/Sentence Ordering:**
  - Is the summary logically organized, does it flow?
- **Fluency:**
  - How grammatically correct is the summary, does it feel natural?
- **Content Selection:**
  - How well is information conveyed without unnecessary repetition, redundancy, and verbosity
- **Overall Quality:**
  - What would you rate it overall?



	ROUGE-1		ROUGE-2	
	Correlation	p-value	Correlation	p-value
Informativeness	0.3484	0.3238	0.0134	0.9706
Coherence	0.3114	0.3810	0.3241	0.3608
Fluency	0.0721	0.8430	-0.3343	0.3450
Content Selection	0.2288	0.5248	0.0635	0.8615
Overall Quality	0.1236	0.7335	-0.2148	0.5511