Do Multi-Document Summarization Models Synthesize?

Presented by:
Kushan Hewapathirana – 229333P

Authors:

1. **Jay DeYoung (h index: 9)**
2. **Stephanie C. Martinez**
3. **Iain Marshall (h index: 24)**
4. **Byron Wallace (h index: 49)**

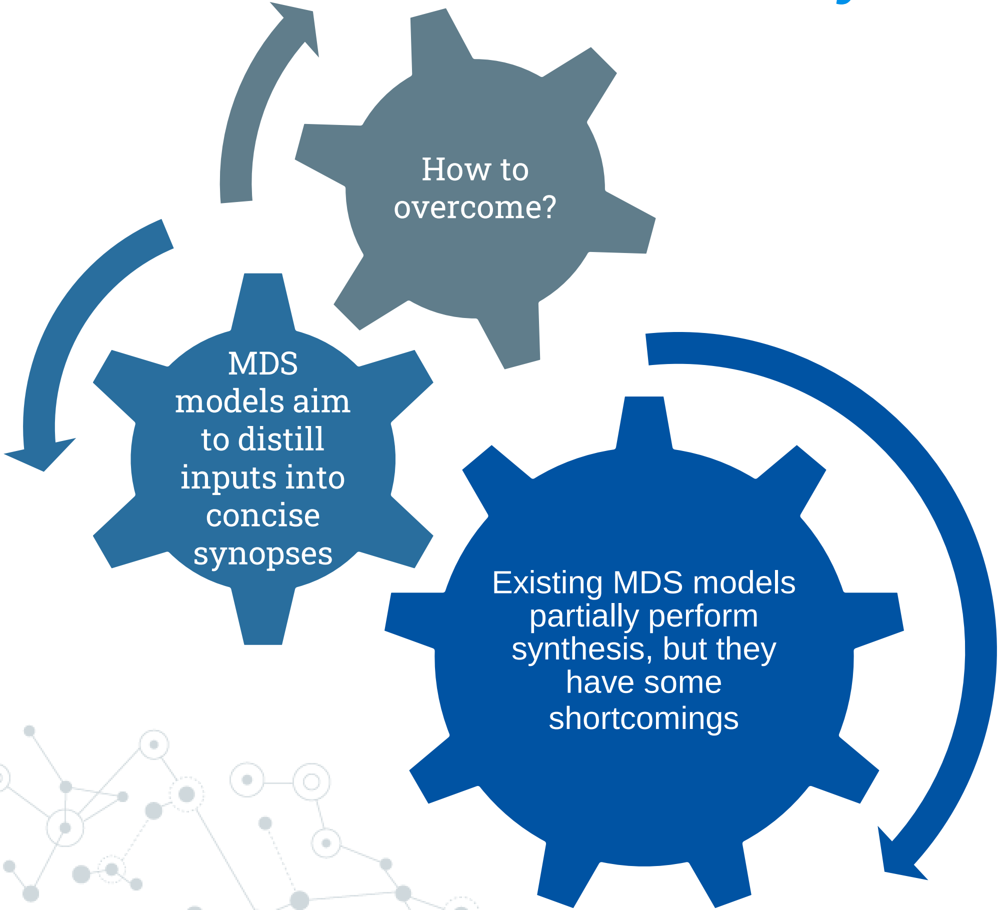
Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA

APPLICATION DOMAIN:

Natural Language Processing – Multi-document Summarization

PROBLEM STATEMENT: Evaluating Synthesis Capabilities of Multi-document Summarization

Challenge against Multi-Document Summarization Models and the Need for Synthesis



How to overcome?

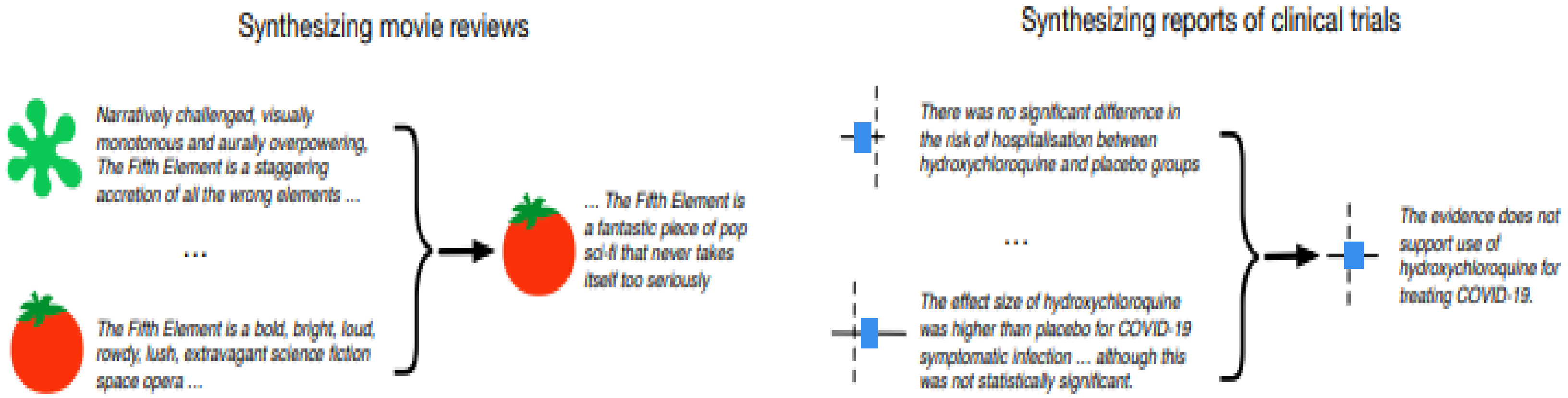
MDS models aim to distill inputs into concise synopses

Existing MDS models partially perform synthesis, but they have some shortcomings

- A movie synopsis should reflect the average opinion of the critics who reviewed it.

- Narrative summaries of biomedical systematic reviews should fairly summarize potentially conflicting results from individual trials.

The Need Of Implicit Synthesis Of Inputs To Produce Accurate Summaries



Synthesizing movie reviews

Narratively challenged, visually monotonous and aurally overpowering, The Fifth Element is a staggering accretion of all the wrong elements ...

...

The Fifth Element is a bold, bright, loud, rowdy, lush, extravagant science fiction space opera ...

... The Fifth Element is a fantastic piece of pop sci-fi that never takes itself too seriously

Synthesizing reports of clinical trials

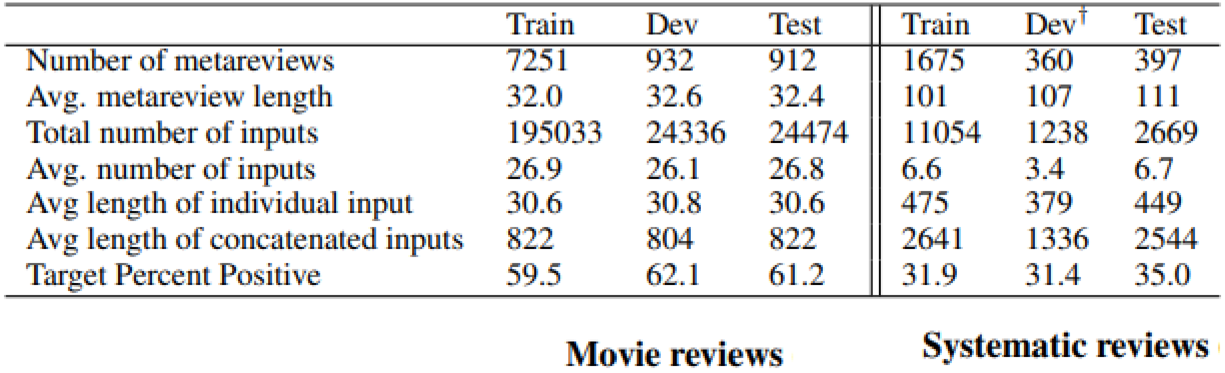
There was no significant difference in the risk of hospitalisation between hydroxychloroquine and placebo groups

...

The effect size of hydroxychloroquine was higher than placebo for COVID-19 symptomatic infection ... although this was not statistically significant.

The evidence does not support use of hydroxychloroquine for treating COVID-19.

Dataset Statistics

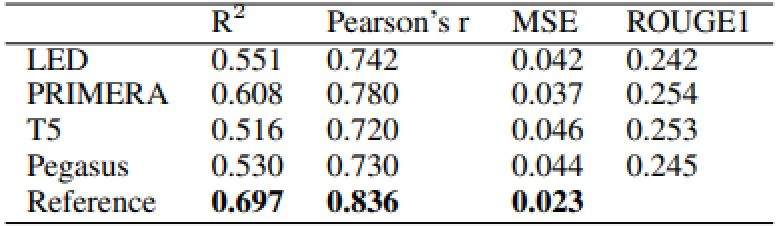


	Train	Dev	Test	Train	Dev†	Test
Number of metareviews	7251	932	912	1675	360	397
Avg. metareview length	32.0	32.6	32.4	101	107	111
Total number of inputs	195033	24336	24474	11054	1238	2669
Avg. number of inputs	26.9	26.1	26.8	6.6	3.4	6.7
Avg length of individual input	30.6	30.8	30.6	475	379	449
Avg length of concatenated inputs	822	804	822	2641	1336	2544
Target Percent Positive	59.5	62.1	61.2	31.9	31.4	35.0

Movie reviews **Systematic reviews**

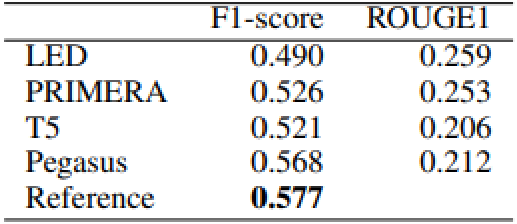
EXPERIMENTS AND RESULTS

How well do summarization models synthesize?



	R^2	Pearson's r	MSE	ROUGE1
LED	0.551	0.742	0.042	0.242
PRIMERA	0.608	0.780	0.037	0.254
T5	0.516	0.720	0.046	0.253
Pegasus	0.530	0.730	0.044	0.245
Reference	**0.697**	**0.836**	**0.023**	

Movie reviews

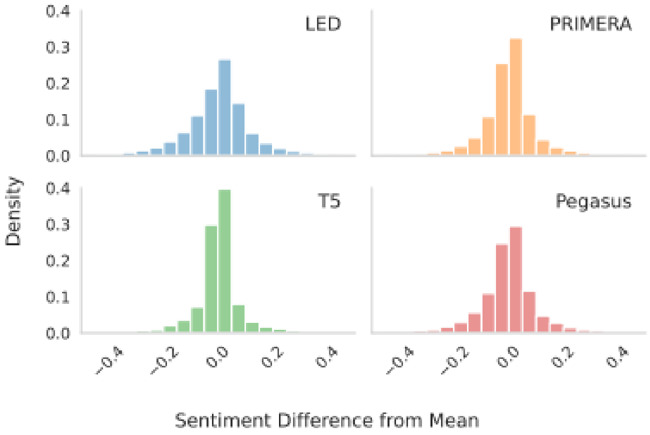


	F1-score	ROUGE1
LED	0.490	0.259
PRIMERA	0.526	0.253
T5	0.521	0.206
Pegasus	0.568	0.212
Reference	**0.577**	

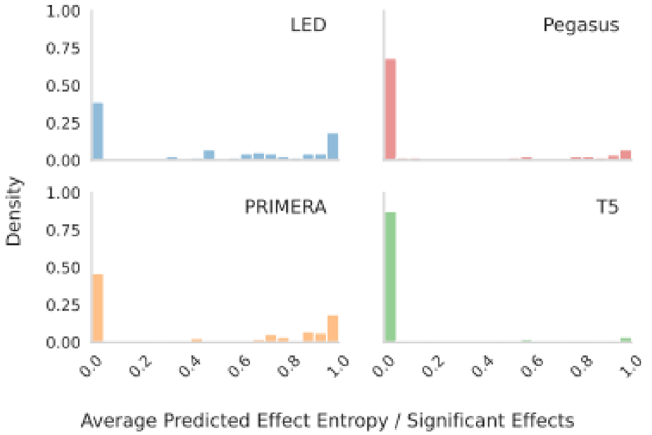
Systematic reviews

- Results suggest that humans perform better in synthesis, as their reported significance in summaries better aligns with the statistical results than in model-generated summaries.

Sensitivity to Input Ordering



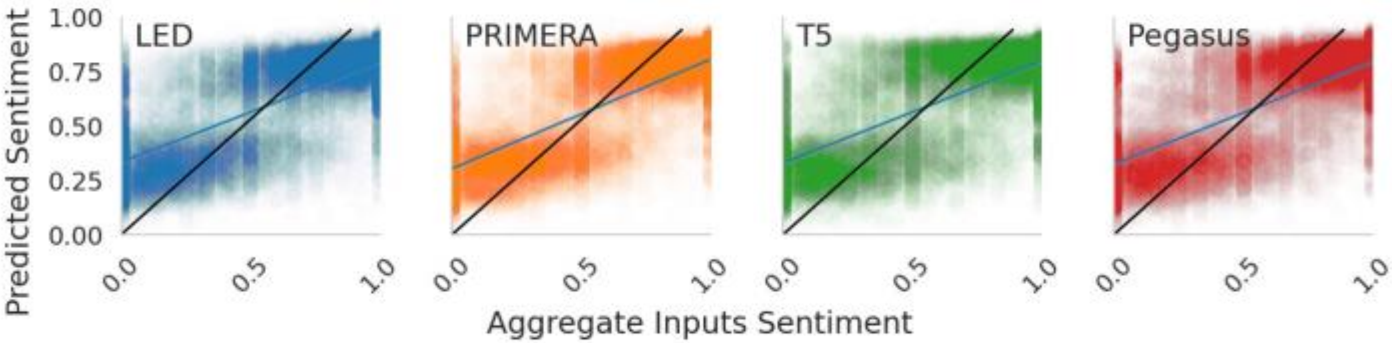
Movie reviews



Systematic reviews

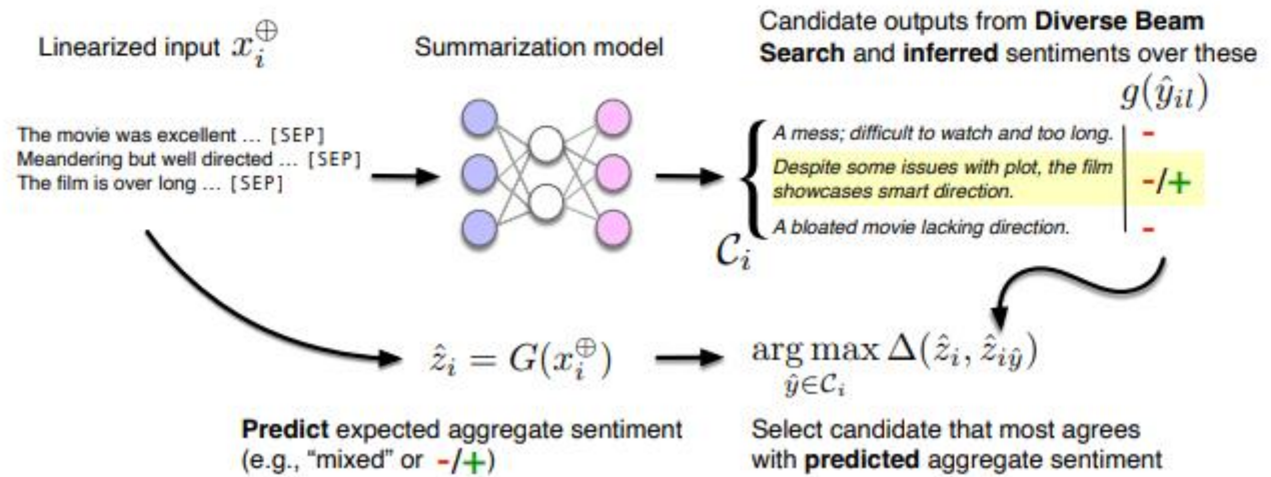
- Synthesis of inputs should be invariant to ordering
- The spread of sentiment/treatment effect measured in outputs produced from permuted input orderings.

Sensitivity to Input Composition



- Synthesis models should be responsive to changes in the distribution of the attribute to be synthesized in the input composition
- The intensity patterns indicate that models tend to oscillate between low and high sentiments in outputs

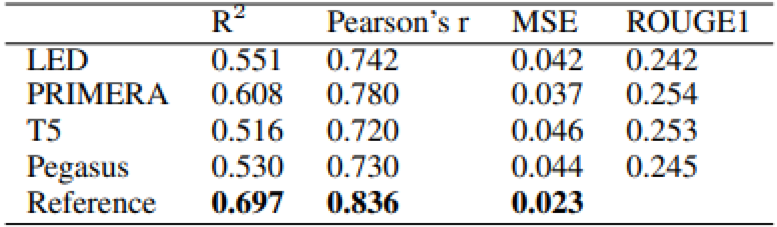
Proposed Strategy To Improve Synthesis


Linearized input $x_i^{\oplus}$
Summarization model
Candidate outputs from **Diverse Beam Search** and **inferred** sentiments over these
$g(\hat{y}_{il})$
The movie was excellent ... [SEP]
Meandering but well directed ... [SEP]
The film is over long ... [SEP]
A mess; difficult to watch and too long.
-
Despite some issues with plot, the film showcases smart direction.
-/+
A bloated movie lacking direction.
-
$\mathcal{C}_i$
$\hat{z}_i = G(x_i^{\oplus})$
$\arg\max_{\hat{y} \in \mathcal{C}_i} \Delta(\hat{z}_i, \hat{z}_{i\hat{y}})$
Predict expected aggregate sentiment (e.g., "mixed" or -/+)
Select candidate that most agrees with **predicted** aggregate sentiment

- Generate an intentionally diverse set of output candidates[1] and then select from these the text that best agrees with the predicted aggregate property of interest

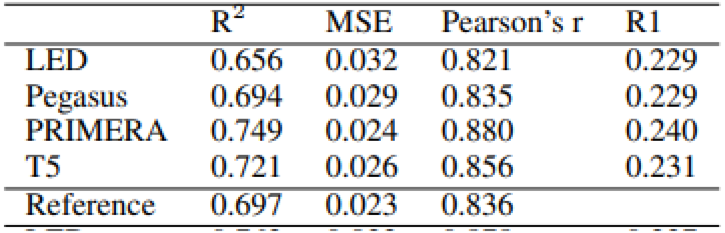
[1] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse Beam Search: Decoding diverse solutions from neural sequence models," *arXiv.org*, 22-Oct-2018. [Online]. Available: https://arxiv.org/abs/1610.02424.

Results



	R^2	Pearson's r	MSE	ROUGE1
LED	0.551	0.742	0.042	0.242
PRIMERA	0.608	0.780	0.037	0.254
T5	0.516	0.720	0.046	0.253
Pegasus	0.530	0.730	0.044	0.245
Reference	**0.697**	**0.836**	**0.023**	

- Without proposed strategy



	R^2	MSE	Pearson's r	R1
LED	0.656	0.032	0.821	0.229
Pegasus	0.694	0.029	0.835	0.229
PRIMERA	0.749	0.024	0.880	0.240
T5	0.721	0.026	0.856	0.231
Reference	0.697	0.023	0.836	

- With proposed strategy

CONCLUSIONS

- Authors have outlined and investigated the problem of synthesis as related to some summarization tasks.

- Existing MDS models partially perform synthesis, but they have some shortcomings

- Authors have proposed and validated a straightforward inference time method to improve model synthesis capabilities by preferentially outputting summary candidates that align with a predicted aggregate measure, and demonstrated empirically that this offers gains in performance.

References

[1] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, “Diverse Beam Search: Decoding diverse solutions from neural sequence models,” *arXiv.org*, 22-Oct-2018. [Online]. Available: https://arxiv.org/abs/1610.02424.

THANK YOU…