

Rationale-based Learning using Self-Supervised Narrative Events for Text Summarisation of Interactive Digital Narratives

Ashwathy T Revi, Stuart E Middleton, David E Millard

Interactive Digital Narratives (IDNs)

- IDNs are narratives that support player interaction.
- Limited research on what prior information about interactive narrative structure can be leveraged for summarization and how



Choices as Rationale for Rationale-Based Learning

- Choices can determine which parts of the narrative are salient for inclusion in the summary
- Rationale based learning using supervised attention has shown good results for text classification
- Using proximity to choice points as a self-supervised proxy for human rationales

Rationales

- Rationales embedded as tensors, indicating proximity to choice points
- Sentence and word level rationales

$$rs_i = \begin{cases} 1 & \text{if } CT \text{ in } (s_{i-ws} : s_{i+ws}) \\ 0 & \text{otherwise} \end{cases}$$

$$rw_i = \begin{cases} tfidf(w_i) & \text{if } w_i \in CW \\ 0 & \text{otherwise} \end{cases}$$

where CW is the set of all words that fall inside a window of size ws around the choice tag given by,

$$CW = \{w_i \in W \mid CT \text{ in } (w_{i-ws} : w_{i+ws})\}$$

CT stands for the choice tag, rs_i and rw_i stand for the rationale for sentence/ word at index i , ws stands for window size, s_i and w_i stands for sentence/ word at index i and notations $s_i : s_j$ and $w_i : w_j$ represents concatenation of sentences/ words at indexes from i to j . In this paper,

Training Objective

For sentence attention model:

$$L = \alpha * L_l + (1 - \alpha) * L_s$$

For word attention model :

$$L = \alpha * L_l + (1 - \alpha) * L_w$$

For attention model with sentence and word level attention :

$$L = \alpha * L_l + \alpha_1 * L_s + \alpha_2 * L_w$$

where: $\alpha + \alpha_1 + \alpha_2 = 1$,

L = Total Loss,

L_l = Cross-entropy loss calculated for the output of the model against the target labels,

L_s = Cross-entropy loss calculated for sentence attention scores against sentence rationales and

L_w = Cross-entropy loss calculated for word attention scores against word rationales.

Experimental SetUp: Dataset

IDNSum - 10,000 documents - playthrough through 8 episodes from two interactive narrative games

Default split (3 episodes from Wolf Among Us is train, remaining 2 validation and 3 episodes from Before the Storm is test set)

Experimental SetUp: Models

SummaRuNNer (RNN) and sentonlyAttnRNN, wordonlyAttnRNN, AttnRNN trained with rationales and without rationales

Longformer and Zero shot Flan T5 - base also shown for comparison.

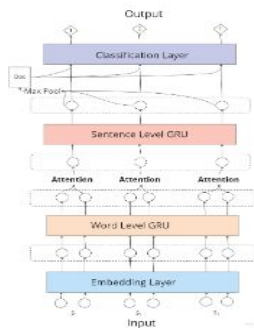


Figure 1: Summarunner modified to use attention instead of max pooling at word level (wordonlyAttnRNN)

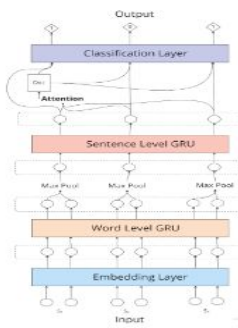


Figure 2: Summarunner modified to use attention instead of max pooling at sentence level (sentonlyAttnRNN).

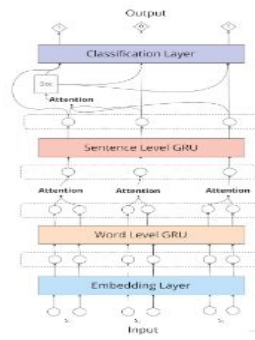


Figure 3: Summarunner modified to use attention instead of max pooling at both word and sentence level (AttnRNN).

Experiment SetUp: Evaluation Strategy

Comparison of Approaches:

1. Similarity with the human written summaries (ROUGE scores)
2. Variability Analysis (Average sentence overlap)
3. Fault Analysis
4. Qualitative Analysis

Results: ROUGE based Automatic Evaluation

Model	R1(abs)	95% CI	R2(abs)	95% CI	RL(abs)	95% CI
SummaRuNNer (RNN)	0.47757	0.47689 - 0.47825	0.12379	0.12323 - 0.124358	0.46460	0.46403 - 0.4651
sentonly AttnRNN	0.44569	0.44464 - 0.44671	0.11624	0.11550 - 0.11697	0.43477	0.43382 - 0.43572
sentonly AttnRNN + rationale	0.50852	0.50767 - 0.50936	0.13036	0.12977 - 0.13095	0.49223	0.49150 - 0.49299
wordonly AttnRNN	0.46508	0.46446 - 0.46568	0.12082	0.12012 - 0.12155	0.45205	0.45152 - 0.45258
wordonly AttnRNN + rationale	0.48124	0.48032 - 0.48209	0.12386	0.12331 - 0.12439	0.46764	0.46681 - 0.46839
AttnRNN	0.44044	0.43983 - 0.44107	0.11081	0.11018 - 0.11142	0.42832	0.42782 - 0.42884
AttnRNN + rationale	0.48637	0.48542 - 0.48725	0.13337	0.13265 - 0.13407	0.47231	0.47147 - 0.47309
Longformer	0.30881	0.30754 - 0.31007	0.06692	0.06641 - 0.06748	0.30237	0.30117 - 0.30354
Google flan-t5-base	0.46577	0.46519 - 0.46637	0.11833	0.11800 - 0.11866	0.41051	0.40997 - 0.41112

Table 1: Mean ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE-L (RL) scores and confidence interval (CI) of generated summaries of IDNSum playthroughs calculated against gold standard human written abstractive summaries(abs).

- Models trained with rationales outperformed those without by up to 14%

Follow up experiments using Flan T5

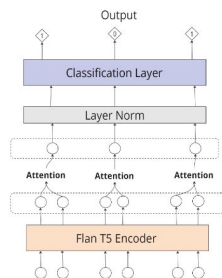


FIGURE 6.4: Flan T5 Encoder with an attention layer and classification head for extractive summarisation.

Google flan-t5-base Encoder	0.44885	0.44794 - 0.44969	0.10322	0.10280 - 0.10364	0.40304	0.40209 - 0.40393
Google flan-t5-base Encoder + rationale	0.47444	0.47398 - 0.47490	0.11660	0.11612 - 0.11707	0.42987	0.42940 - 0.43031

Results: Variability Analysis

Model	Avg overlap
RNN	47.85
sentonly Attn	53.48
sentonly Attn + rationale	44.76
wordonly Attn	50.84
wordonly Attn + rationale	49.66
AttnRNN	49.21
AttnRNN + rationale	45.88

Table 2: Average number of overlapping sentences for every pair of summaries from each episode for each model (out of a total of 81 sentences).

- Summaries produced by rationale-based models showed up to 16% more variability across playthroughs

Results: Fault Analysis

- Four main error types: Redundancy, Incompleteness, Irrelevance, and Unclear Sentences
- Redundant, Incomplete and irrelevant information were the most common errors

Error Type	Ep 1	Ep 2	Ep 3	Avg
Redundant	16.5	13.9	22.3	17.57
Incomplete	18.9	17.4	13.9	16.73
Irrelevant	15.2	17.4	21.5	18.03
Unclear	0.1	0.5	0.1	0.23

Table 3: Fault Analysis : Error types in model generated summaries and the average number of sentences exhibiting these errors out of a total of 81 sentences per summary.

Results: Qualitative Analysis

- Fragment view of the narrative.
- Sometimes context is missing and some inference is required.
- Could potentially serve as recap.
- Coverage of choices and differences across playthroughs remains a challenge

"chloe price, standing on train tracks and wearing a black hoodie, flicks her lighter a few times and lights up her cigarette. a train begins to approach her. after a few moments, the guy she ran into earlier and his friend come to confront her. rachel takes chloe's hand again and they run towards the entrance to the show. frank sees them and chloe stops, looking at the guys behind him. the men leave and frank looks back to see that rachel and chloe are gone. if she attacked the skeevy guys, she will now have a bruise under her eye."

Limitations and Future Work

- Fault analysis done by a single annotator, introducing potential subjectivity
- Results shown are for single runs
- Only one type of IDN, limiting generalizability. Future work includes evaluating this approach on diverse datasets and architectures

Conclusion

- Choices can act as effective explanations for IDN summarisation, improving performance up to 14% and variability up to 16%.
- Sentence level rationales performed best.
- The manual analyses show limitations of the approach.
- Encourages rationale based learning for other types of narratives.

Thank you!

atr1n17@soton.ac.uk

<https://www.linkedin.com/in/ashwathytr/>