SummScreen: A Dataset for Abstractive Screenplay Summarization

Mingda Chen[1], Zewei Chu[2], Sam Wiseman[3], Kevin Gimpel[1]

[1]Toyota Technological Institute at Chicago   [2]University of Chicago   [3]Duke University

SummScreen

- An abstractive summarization dataset combining TV series transcripts and episode recaps.
- We collected data from fan-contributed websites.
- We divide SummScreen into two subsets, FD and TMS, according to data sources.

	FD	TMS
#shows	88	10
#episodes	4348	22503
#lines in trans.	447.6	360.8
#char. utterances in trans.	330.7	327.0

Transcript from "The Big Bang Theory":

[The apartment]

Sheldon : What color would you like to be ?
Leonard : Well , I 'd like to be green , but you know you always take it .
Sheldon : That 's not true . Any color 's fine with me . Yeah , I could be a - a combination of blue and yellow .
Leonard : Blue and yellow make green .
Sheldon : Well , then it 's settled

Penny : Hi . Ready to go ?
Sheldon : Oh , good news , we ordered lunch , so we can all stay here and play Lord of the Rings Risk .
Amy : Sheldon , we said that we would play games with you tonight .
Sheldon : Oh , no , we 'll still be playing it tonight , this game can easily take eight hours .
Penny : Sweetie , you really thought I 'd want to do this ?

Leonard : No .
Penny : Well , did you tell him that ?
Leonard : Yes .
Penny : Did you say it out loud with words ?
Leonard : No .
Penny : I do n't want to spend the whole day playing a board game .

…

Human-Written Recap:

Sheldon and Leonard are happy playing a board game until Amy and Penny say they are tired of doing what the guys want …

☐ Colored boxes indicate utterances belonging to the same conversations.

— Clues in the transcript are underlined.

Dataset Comparison

	#inst	len(inp)	len(outp)	#speak.	Domain
SAMSum	16.4k	83.9	20.3	2.2	ChitChat
ForumSum	4.1k	303.5	36.0	6.7	Forum msg.
MediaSum	463.6k	1553.7	14.4	6.5	Interviews
AMI	137	4757.0	322.0	4.0	Meetings
ICSI	59	10189.0	534.0	6.2	Meetings
QMSum	1.8k	9069.8	69.6	9.2	Meetings
Our work	26.9k	6612.5	337.4	28.3	TV series

- SummScreen combines long source inputs, large numbers of speakers, and a moderate number of instances.

Dataset Challenges

- Understand the context in which characters' utterances are situated.
- Draw information from a wide range of the input transcripts.

Transcript from "Doctor Who":

Line 119 DOCTOR : Camera ! Camera ! (takes camera from ALEC 'S unresisting hands)

…

Line 212 The DOCTOR turns around and continues to take photos with the camera …

…

Line 287 The DOCTOR steps out of the TARDIS wearing the spacesuit … He scans with the sonic before picking up the camera to take a few pictures .

…

Line 336 DOCTOR : Right ! Done ! That 's it … She 's not a ghost ... but she 's definitely a lost soul . (walks over to screen) Her name 's Hila Tacorian . She 's a pioneer , a time traveller - or at least she will be , in a few hundred years .

Human-Written Recap:

… the Doctor borrows Alec 's camera and uses the TARDIS to take pictures of the mansion 's location throughout time . Thanks to this , the Doctor learns it 's not a ghost in the pictures , but a time traveler named Hila Tacorian …

Line xxx: line numbers in the original transcript.

Experiments

Approaches

- Neural models
- Nearest neighbor models (NNM): retrieve summaries from the training set.
- Hybrid models (nearest neighbor content selector → neural models)

Metrics

- BLEU and ROUGE
- Entity metrics:
  - Bag of characters (BoC): the fraction of the **characters** overlapping with gold
  - Bag of character relations (BoR): the fraction of the **cooccurred character pairs** overlapping with gold

Results

	BLEU	R1	R2	RL	avg
NNM	7.9	31.3	7.8	27.4	18.6
Neural Model	2.6	25.9	4.2	23.8	14.1
Hybrid Model	2.4	25.3	3.9	23.1	13.7
Hybrid Model (w/ oracle content selector)	3.0	26.4	5.0	23.3	14.4

	BoC-p	BoC-r	BoR-p	BoR-r	avg
NNM	56.7	59.2	28.2	29.4	43.3
Neural Model	54.7	38.5	22.8	15.1	32.8
Hybrid Model	61.2	51.4	29.8	23.6	41.5
Hybrid Model (w/ oracle content selector)	70.0	57.8	36.9	29.1	48.5

FD Test Results

- NNM shows strong performance.
- With the help of the oracle content selector, the hybrid model improves significantly in both semantic matching and entity-related metrics.