Seinfeld TV Script Generator

Yueyan Lu, Miaozhi Yu

Executive summary

- **Goal**: Design and train deep neural networks that generate TV scripts for the show Seinfeld. It should be able to
  - generate TV script for individual character.
  - generate dialogue between two characters that frequently chat with each other.
- **Approach**: Implement several Recurrent Neural Networks (RNN) models or Seq2seq models and evaluate their corresponding performance.
- **Value/Benefit:** With the TV script generator, we hope to reduce the human effort in writing the day-to-day conversation between main characters.

Problem motivation

- With the advances in deep learning and neural network, a lot of research has been conducted in text generation. However human dialogue generated by the neural network often suffers from incomprehension and incoherence to a human reader.
- In this project we would like to explore how different models can help improve the coherence in TV script dialogue generation.
- *Seinfeld* is an American sitcom aired between 1989 to 1998, with four major characters, each of whom has distinctive personality. We would also like to explore if the model can successfully capture the characteristics of each role.

Background work

- **Paper**:
  - Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.
- **GitHub Repo**:
  - Natural Language generation using Neural Machine Translation Context (https://github.com/samjkwong/NLG-NMT)
- **Technical blogs**:
  - https://www.geeksforgeeks.org/ml-text-generation-using-gated-recurrent-unit-networks/
  - https://medium.com/@david.campion/text-generation-using-bidirectional-lstm-and-doc2vec-models-2-3-f0fc07ee7b30
  - https://towardsdatascience.com/perplexity-in-language-models-87a196019a94

Technical challenges

- Raw Data Processing:
  - Large Data Set: 9 Seasons 180 episodes (39K lines for main characters)
  - Vocabulary building: per character / word embedding including dealing with special characters
- Model Selection:
  - Character-based GRU
  - Word-based Bidirectional LSTM
  - Character-based NMT
- Evaluation Metrics Selection:
  - BLEU
  - Perplexity
  - Human judgement on properties like sentence coherence and humour etc
- Generate sentences with sensible dialogue:
  - Proper sentence structure for individual lines
  - Capture general sentiment of previous text in dialogues

Approach-GRU

Model Summary:

Model: "sequential_6"

Layer (type)	Output Shape	Param #
gru_6 (GRU)	(None, 128)	83712
dense_6 (Dense)	(None, 88)	11352
activation_6 (Activation)	(None, 88)	0

Total params: 95,064
Trainable params: 95,064

- Trained character-based GRU model for each main character
  - Used RMSprop optimizer
  - Learning rate with decay schedule
- However, the performance is very poor. It sometimes is not able to generate correct english words.
  - For example:
    jerry: yeirfies.
    jerry: you've hichow
- Even if the character-based model captures the word structure, it is not able to generate meaningful sentences
  - For example:
    jerry: with all do respect for this my face any

Approach-LSTM

Model Summary:

Layer (type)	Output Shape	Param #
bidirectional_4 (Bidirection	(None, 512)	44347392
dropout_4 (Dropout)	(None, 512)	0
dense_4 (Dense)	(None, 21397)	10976661
activation_4 (Activation)	(None, 21397)	0

Total params: 55,324,053
Trainable params: 55,324,053
Non-trainable params: 0

Total sequences: 236157

- To address the incomprehension issue experienced in the GRU model, we trained word-based bi-directional LSTM model on each individual character and collectively on all four leading roles.
- With the LSTM model, most of the time the predicted sentences has proper sentence structure.

```
jerry: can you relax, its a cup of coffee. claire is a professional waitress.

jerry: well, theres this uh, woman might be comin in.
```

- However, we found the LSTM model still failed to capture the general sentiment of previous text in dialogue generation, leading to unrelated lines

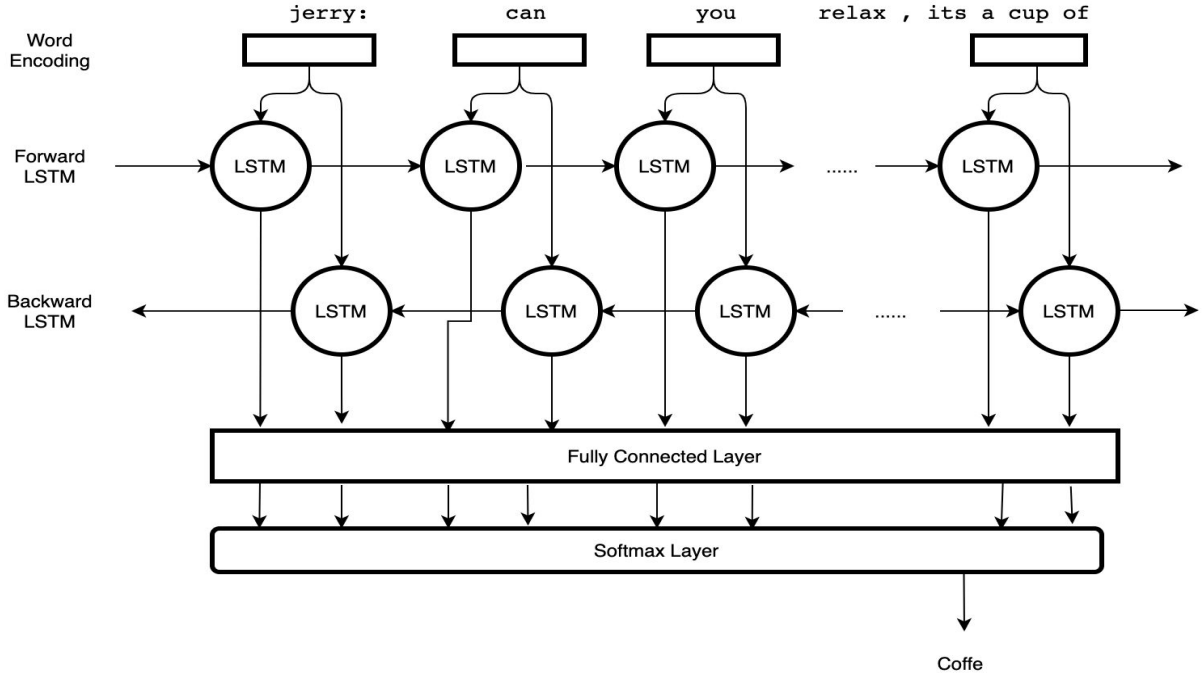
```
jerry: i don't know what you do.

george: i can't believe this guy.

jerry: i can't believe this guy. i can't believe he hasn't called me to see.
```

Approach-NMT

- LSTM models are able to generate somewhat sensible text based on the lines of each main character. However, it is not enough for TV script as it is not able to generate sensible dialogue between characters
- Explored NMT
  - Model Architect:
    - First used character-based convolutional encoder to derive word embedding
    - Then an RNN to encode line of dialogue as a sequence of encoder hidden states
    - Another RNN to produce decoder hidden states based on target sentence using attention on encoder hidden states
    - Finally, a softmax layer to predict the most likely target word
  - Training:
    - Trained NMT for each main character
    - Used the character's line as target sentence and the line before the character's line as source sentence

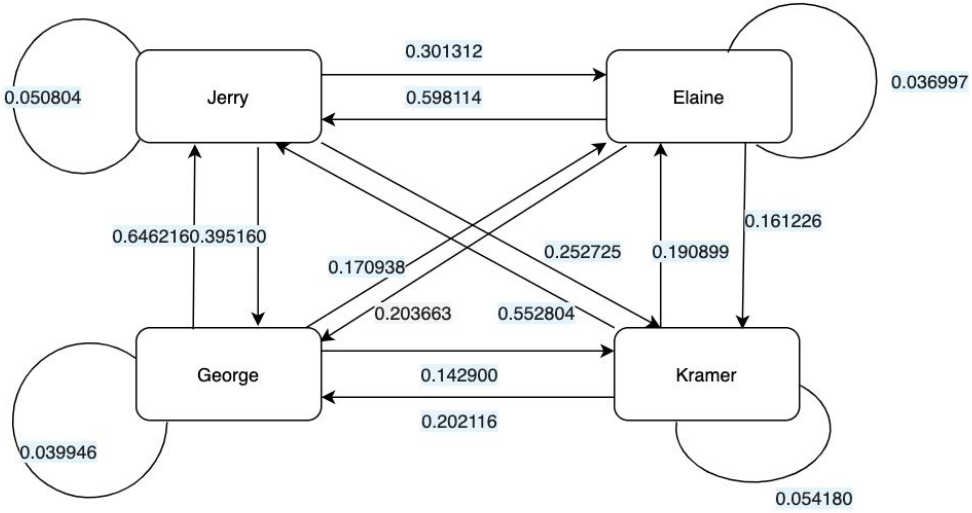
Solution diagram/architecture

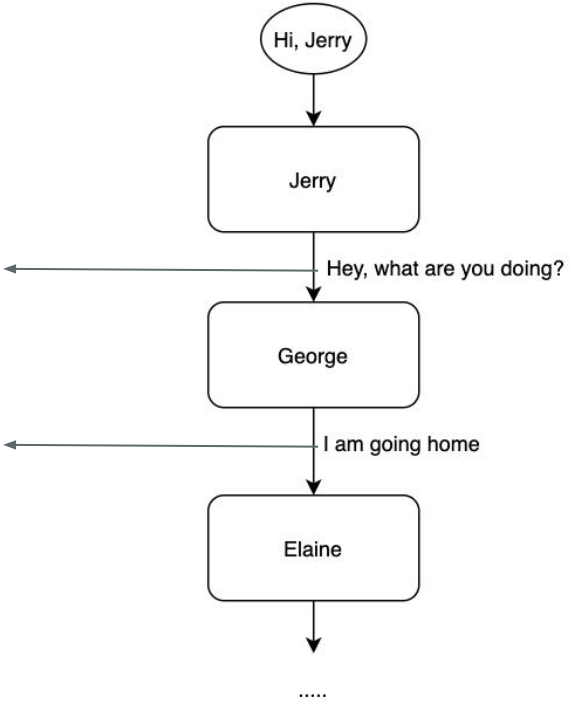
LSTM Model Structure



Solution diagram/architecture

NMT Solution Structure





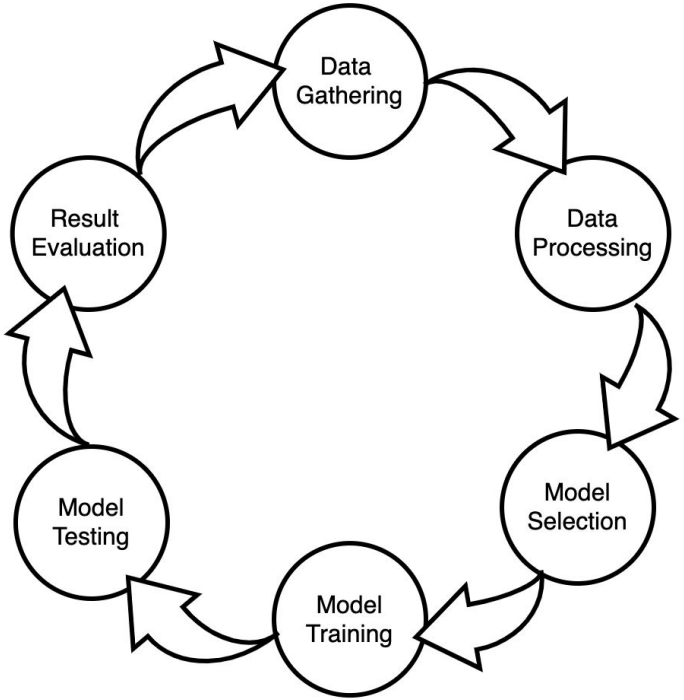
Implementation details

- Dataset

	Character	Dialogue	EpisodeNo	SEID	Season	char_dial
0	JERRY	Do you know what this is all about? Do you kno...	1.0	S01E01	1.0	jerry: Do you know what this is all about? Do ...
1	JERRY	(pointing at Georges shirt) See, to me, that b...	1.0	S01E01	1.0	jerry: (pointing at Georges shirt) See, to me,...
2	GEORGE	Are you through?	1.0	S01E01	1.0	george: Are you through?
3	JERRY	You do of course try on, when you buy?	1.0	S01E01	1.0	jerry: You do of course try on, when you buy?
4	GEORGE	Yes, it was purple, I liked it, I dont actuall...	1.0	S01E01	1.0	george: Yes, it was purple, I liked it, I dont...

- LSTM & GRU:
  - Keras with Tensorflow backend
  - GCP V100
- NMT
  - Pytorch
  - GCP K80

Demo/Experiment design flow



- Phase 1: Character-based GRU Model
- Phase 2: Word-based bi-directional LSTM Model
- Phase 3: NMT Model with character-base convolutional encoder

Experimental evaluation

Model: GRU

Generated Text:

jerry: What do you mean you want to get the same thing is the bottom. You know what you think that doesn't make it out of….

george: Well, I don't know what they want to the postrouse and the many to the point. I can't see that me.

elaine: (exciter) You're the connation of the stopped me.
elaine: Well, I can't

Evaluation Metrics

BLEU:

Cumulative 1-gram: 0.2624819812758244
Cumulative 4-gram: 0.03859844449749971

Perplexity: 4.482171403098487e-05

Experimental evaluation

Model: Bi-directional LSTM

Generated Dialogue:

jerry: i don't know what you do.
george: i can't believe this guy.
jerry: i can't believe this guy. i can't believe he hasn't called me to see.
elaine: what are you doing?
jerry: i don't know.
elaine: what?
jerry: i don't know.
elaine: you know, i think this is a little awkward.
elaine: what is this?

Evaluation Metrics

BLEU:

Cumulative 1-gram: 0.45539999999999997
Cumulative 4-gram: 0.1934442068434737

Perplexity: 0.00019151019468558275

Experimental evaluation

Model: NMT

Generated Dialogue:

ELAINE: Hi.
GEORGE: Kramer?
ELAINE: I know!
GEORGE: All right, I think I may have to go back back.
ELAINE: She said you're getting together Saturday night!
GEORGE: Oh no no no no no no no.
ELAINE: Yeah!
GEORGE: I think I don't understand.

Evaluation Metrics

BLEU:
1.5753910242882029e-77

Perplexity:
9907.62

Experimental evaluation

Comparison across all the models:

	BLEU	Perplexity
GRU	0.262481981275824	4.48217140309848e-05
LSTM	0.3580000000000004	0.00019151019468558275
NMT	1.57539102428 8e-77	9907.62

Note:

1.BLEU (etc) scores are only valuable if they can reliably predict the result of testing the hypotheses we care about (utility, etc).

2.Perplexity is still useful for comparing our baseline language models to each other, since they have similar structures in architecture, and we could empirically compare how good they were at predicting the next word.

3. We also would like to use human judgement to evaluate the quality of the dialogues

Conclusion

- In this project, we explored three different models in producing plausible Seinfeld scripts and evaluate the model performance accordingly.
- Although we did not achieve state-of-art result for the script dialogue generation, the model we experimented meets the basic requirements of generating sentences with proper structure and context.
- As next step of improving the coherence between back-and-forth conversation, we suggest bring in analysis on the text topic using methods like LDA to improve the semantic relation in the dialogue and overall quality of the script.
- Explore other academic work on related topic.

Thank you!
Question?

Links

Github

- https://github.com/joyceluyy/Deep_Learning_System_Project

Readme