Harry BOTter and the Order of the N-gram

Krithika Iyer, Kalyani Ramadurgam, Neha Sharma

{ksiyer, kalyrama, nesharma}@stanford.edu

Introduction

This project aims to create a chatbot that allows users to interact with their favorite fictional characters and generates new dialogue between characters. We focus on the Harry Potter books and movies, for which dialogue generation is difficult due to the small corpus per character. To solve this, we use a twofold approach: first, we conduct analysis on character dialogue using Jaccard similarity. Next, we use these insights to pre-process the data and combine the training corpus for similar characters. Second, we use an LSTM to generate character dialogue based on a chosen initial seed, and we compare this performance to a GPT model fine-tuned on Harry Potter dialogue and books.

Data

- Combined corpus of all Harry Potter movie dialogue and book text
- Data corpus:

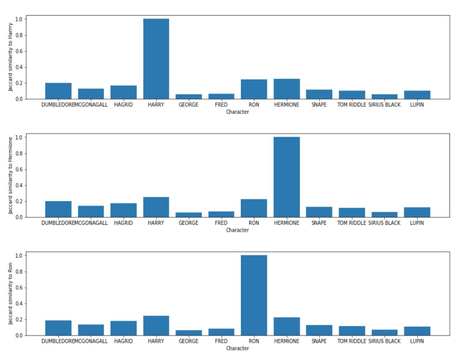
```
DUMBLEDORE 312
MCGONAGALL 51
HAGRID 206
HARRY 1387
GEORGE 50
FRED 61
RON 715
```

```
HERMIONE 703
SNAPE 103
PROFESSOR MCGONAGALL 61
TOM RIDDLE 70
PROFESSOR LUPIN 86
SIRIUS BLACK 36
LUPIN 18
```

```
ALL BOOKS COMBINED:
for all combined:
Total Tokens: 1003024
Unique Tokens: 21512

ALL Dialogs: (after tokenization)
Total Tokens: 40195
Unique Tokens: 4332
```

- Initially used Jaccard Similarity to score similarity of character dialogue and combine characters to increase training data volume. This has three steps.
  - Remove extraneous punctuation (periods, commas, ellipses)
  - Tokenize and merge book/movie dictionaries of character dialogue
  - Calculate pairwise Jaccard similarity between tokenized sets of characters
- Jaccard similarity of all characters to Harry, Hermione, and Ron:



LSTM Model

- Created two basic LSTM word and character-based language models
- Model was trained using quotes from movie and book dialogue to generate quotes for a particular character in the Harry Potter series
- Basic model consists of two LSTM hidden layers with 200 and 100 memory cells. Configuration includes a drop-out layer. Length of word sequence is 50. A soft-max output activation function is used. Configuration of baseline model shown below:

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 50, 50)	89050
lstm_1 (LSTM)	(None, 50, 200)	200800
lstm_2 (LSTM)	(None, 100)	120400
dense_1 (Dense)	(None, 100)	10100
dropout_1 (Dropout)	(None, 100)	0
dense_2 (Dense)	(None, 1781)	179881

Total params: 600,231
Trainable params: 600,231
Non-trainable params: 0

- BLEU produced meaningless results, as the purpose was creative text generation, not repetition of existing phrases

GPT Model

	Layers	Required Disk space	Credibility Score (from Cornell) out of 10	Finetuability
117M - Small	12	500M	NA	Easier to finetune
345M - Medium	24	1.5 GB	6.07	Easy
763M - Large	36	3.0 GB	6.72	More difficult
1542M - XL	48		6.91	Can not be finetuned

- Fine-tuned “small-size” model with dialog data and “medium-size” model with book data

Acknowledgements

We want to thank the entire CS221 teaching staff, our TA mentor, Chuma Kabaghe, as well as Professor Percy Liang and Professor Dorsa Sadigh. Additionally, we would like to thank Google Collaboratory for access to a GPU.

Results & Discussion

- Sample LSTM Output:
  - Seed: *who owned that wand he killed my parents didn’t he*
  - Output: *monster at the back door did you find my parents*
- Sample GPT Output:

```
Harry, Ron, and Hermione were not alone, but he was
sure that there were other people like them, and now that
he had found them, they would be able to do as they pleased. . . .
Harry had just finished rummaging around in the lockers when
he heard what sounded like his mother’s voice.
“You’ll be okay,” she had said, in a voice that sounded
almost as though she had grown still taller. She
was staring at him, looking very worried.
“Oh, you’ll be alright,” said Harry, as he placed
his hand on the back of her head. “I’ll call Dad
and see what he can do. . . .”
```

```
DUMBLEDORE:
Well?

HARRY:
Just tell me.  What was that like?

DUMBLEDORE:
I wasn’t there, was I?  Just
said, if I could’ve known.  Ask
anyone.
```

- Output validation (anonymous survey of 60 students):
  - Shown 4 GPT-generated texts (were not told if they were computer generated or portion of a real fanfiction) and asked to rate whether they were realistic or not
  - 73% of respondents marked the 4 texts as “realistic”
  - 35% of respondents had read Harry Potter fanfiction before
    - Of these respondents, 52% of respondents marked the 4 texts as “realistic”

Conclusion & Future Work

- Showed that the GPT model is robust in the face of low training data size for text generation when compared to LSTM
- Moving forward, allow users to alter seed text in order to simulate conversation with fictional character
- Possible to study the ease with which fake news can be generated through the GPT HP model (make characters say OOC dialogue)

References

[1] https://medium.com/huggingface/how-to-build-a-state-of-the-art-conversational-ai-with-transfer-learning-2d818ac26313
[2] https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/
[3] http://jalammar.github.io/illustrated-transformer/
[4] https://deepai.org/machine-learning-glossary-and-terms/jaccard-index
[5] https://skymind.ai/wiki/bagofwords-tf-idf
[6] https://openai.com/blog/gpt-2-1-5b-release/
[7] https://d4mucfpksywv.cloudfront.net/papers/GPT_2_Report.pdf
[8] https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf
[9] http://karpathy.github.io/2015/05/21/rnn-effectiveness/
[10] https://github.com/minimaxir/gpt-2-simple