Information retrieval based south park chatbot

Lingyang Ji, Jianqiao Liu, Yuxin Xie, Zexin Liao

1.Introduction

Our team built a “South Park Chatbot” based on information retrieval methods and dataset from Kaggle website. Creating a dialogue system by retrieving proper cartoon lines is a good way to test our methods. Therefore, we implemented four methods, which are TF-IDF, language model, Convolutional Neural Network and TF-IDF 3, to build a chatbot. Following are the detailed descriptions of our approaches, experimental setup which includes evaluation and results, and possible future work which were discussed by our teammates and communicated with you during presentation process.

2. Approach

First, we applied data visualization to analyze the data and used NLTK python package to conduct word stemming and stop words removal. Next, we split the dataset with the ratio of 8:1:1 as the training, validation and testing datasets. Finally, we proceeded to the following three approaches:

2.1 TF-IDF based method.

The first method is to obtain the sentence with the highest cosine similarity using TF-IDF representation in training data with each sentence in test data. Then the response of this sentence, which is the next sentence, is the answer.

2.2 Language model with Laplace smoothing based method.

The second method is to apply the language model with Laplace smoothing to score the sentences and the other steps are the same as the first method.

2.3 Convolutional neural networks based method.

The third approach is to use convolutional neural networks to learn the similarity between two sentences in the training dataset. If the two sentences are continuous, we labeled them as 1, if not, we labeled them as 0. Then we used this model to choose the sentences from the training dataset with the best similarity score for the sentence in the test dataset.

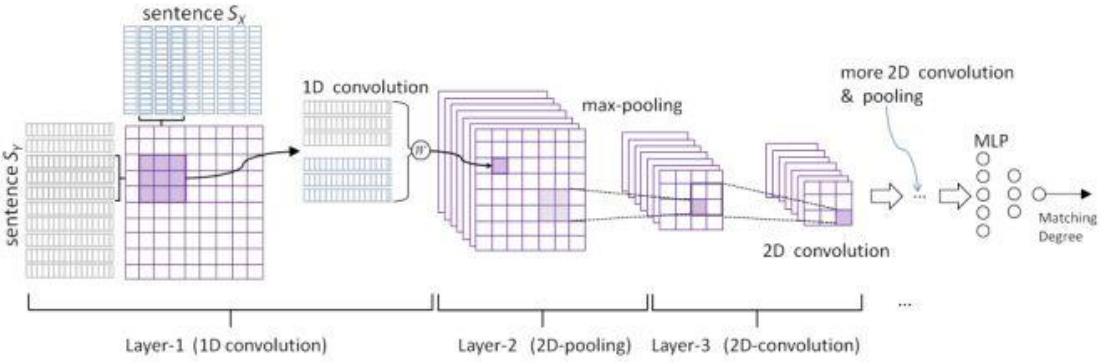


Figure 1: CNN model

Figure 1 shows our 4-layer CNN model. The first layer is the 1D convolution layer and we obtained matrices with two sentences. The second layer is the 2D pooling layer and we used max-pooling to compress the information. The third layer is the 2D convolution layer. The last layer is the fully connected layer and the output is one score of the similarity.

2.4 TF-IDF 3 based method

The last method was similar to the first method. However, we added two more sentences to score the similarity. Then we combined three continue sentences to predict the fourth sentences with weight 0.1, 0.2, 0.7. The latter the sentence, the larger the weight.

3.Experimental Setup

In this section, we will describe our dataset, evaluation method, and experimental result.

3.1 Datasets

The dataset we use is downloaded from Kaggle Website. It includes all the lines of characters in South Park anime across nine seasons and hundreds of episodes. There are 70,897 lines in total. From left to right, features in each line are line number, season number, episode number, character name, character’s line. In this experiment, we will ignore characters’ individual personality and take all the dialogue as a style of “South Park Style”, so the only features we will keep is line number and the sentence itself.

3.2 Evaluation

The performance of the dialogue system will be evaluated in the test set in 4 ways by two BLEU n-gram parameters and two human evaluation standards.

3.2.1 BLEU-n

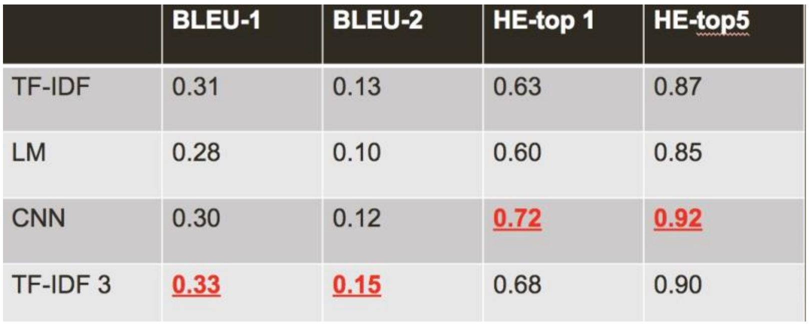
BLEU is bilingual evaluation understudy, which is a popular algorithm for evaluating the quality or similarity of text. It gives the sentence a score using “n-grams”, which check n contiguous sequence of the word in the sentence. At BLEU-1, we check the sentence word one by one and then divided by the total word amount. At BLEU-2, we check contiguous word group two by two.

3.2.2 HE-top n

HE represents “Human Evaluation”. We manually go over 400 query results. At HE-top 1, if the top 1 response seems to be a reasonable response, we will give it score 1. Otherwise, it will get score 0. At HE-top 5, if one of the responses in the top 5 response makes sense, it will get score 1. Otherwise, it will get 0.

3.3 Results

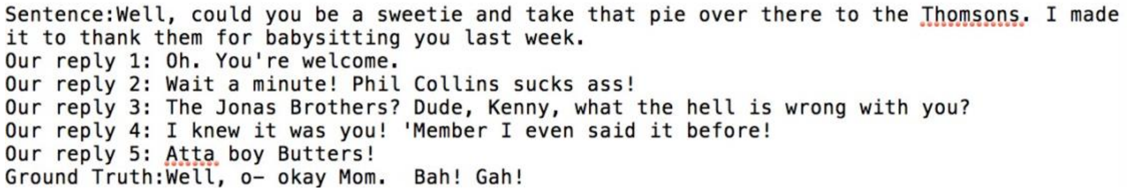
According to the evaluation methods, the results can be calculated as the following form.



	BLEU-1	BLEU-2	HE-top 1	HE-top5
TF-IDF	0.31	0.13	0.63	0.87
LM	0.28	0.10	0.60	0.85
CNN	0.30	0.12	**0.72**	**0.92**
TF-IDF 3	**0.33**	**0.15**	0.68	0.90

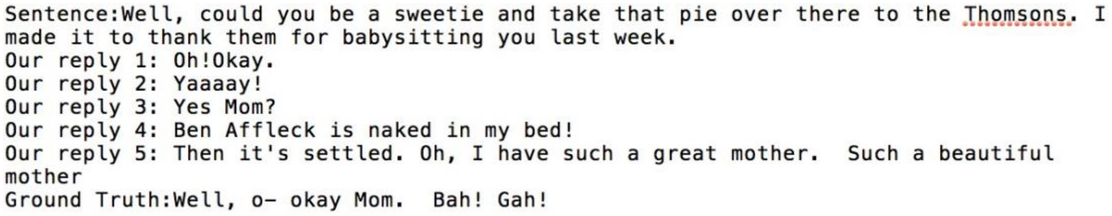
Table 1: Results

As can be seen in table 1, CNN and TF-IDF 3 have the best performance for human evaluation and BLEU, and the corresponding scores are highlighted in red. Specifically, typical output examples containing top 5 replies and ground truth of CNN, TF-IDF 3 and baseline TF-IDF are shown in the following figures.



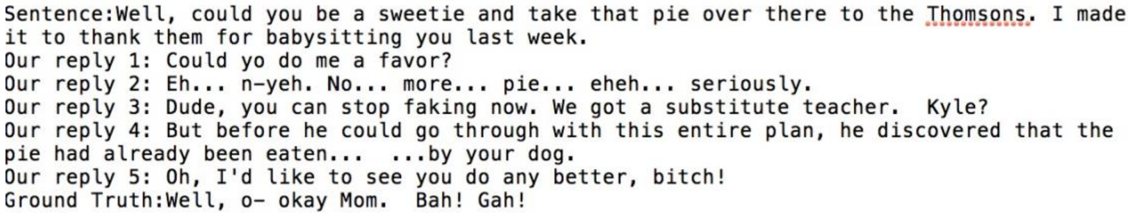
```
Sentence:Well, could you be a sweetie and take that pie over there to the Thomsons. I made
it to thank them for babysitting you last week.
Our reply 1: Oh. You're welcome.
Our reply 2: Wait a minute! Phil Collins sucks ass!
Our reply 3: The Jonas Brothers? Dude, Kenny, what the hell is wrong with you?
Our reply 4: I knew it was you! 'Member I even said it before!
Our reply 5: Atta boy Butters!
Ground Truth:Well, o- okay Mom.  Bah! Gah!
```

Figure 2: CNN



```
Sentence:Well, could you be a sweetie and take that pie over there to the Thomsons. I
made it to thank them for babysitting you last week.
Our reply 1: Oh!Okay.
Our reply 2: Yaaaay!
Our reply 3: Yes Mom?
Our reply 4: Ben Affleck is naked in my bed!
Our reply 5: Then it's settled. Oh, I have such a great mother.  Such a beautiful
mother
Ground Truth:Well, o- okay Mom.  Bah! Gah!
```

Figure 3: TF-IDF 3



```
Sentence:Well, could you be a sweetie and take that pie over there to the Thomsons. I made
it to thank them for babysitting you last week.
Our reply 1: Could yo do me a favor?
Our reply 2: Eh... n-yeh. No... more... pie... eheh... seriously.
Our reply 3: Dude, you can stop faking now. We got a substitute teacher.  Kyle?
Our reply 4: But before he could go through with this entire plan, he discovered that the
pie had already been eaten...  ...by your dog.
Our reply 5: Oh, I'd like to see you do any better, bitch!
Ground Truth:Well, o- okay Mom.  Bah! Gah!
```

Figure 4: TF-IDF(Baseline)

According to above figures, the top 5 replies are reasonable and CNN and TF-IDF 3 generally perform better than TF-IDF.

4.Future work

After discussion with the professor during the final presentation, we figured out three possible methods that may improve the performance of our chatbot. First, because the TF-IDF 3 result performed better than traditional TF-IDF method, we still would like to try to implement other methods to improve the TF-IDF way. We may try to choose top k (k may be 100) sentences with the highest score, then use CNN to rank again in order to save the ranking time. Second, we also want to try LSTM to build multi-round dialogue information to score the similarity of sentences. Third, we may add weight to both of the results of TF-IDF and CNN methods. Then we will reconsider the performance by the weight and regenerate the possible results.