

Basketball Chatbot Report

Kevin Shi & Hrithik Choudhary

CS 4395.001

Mazidi

System Description

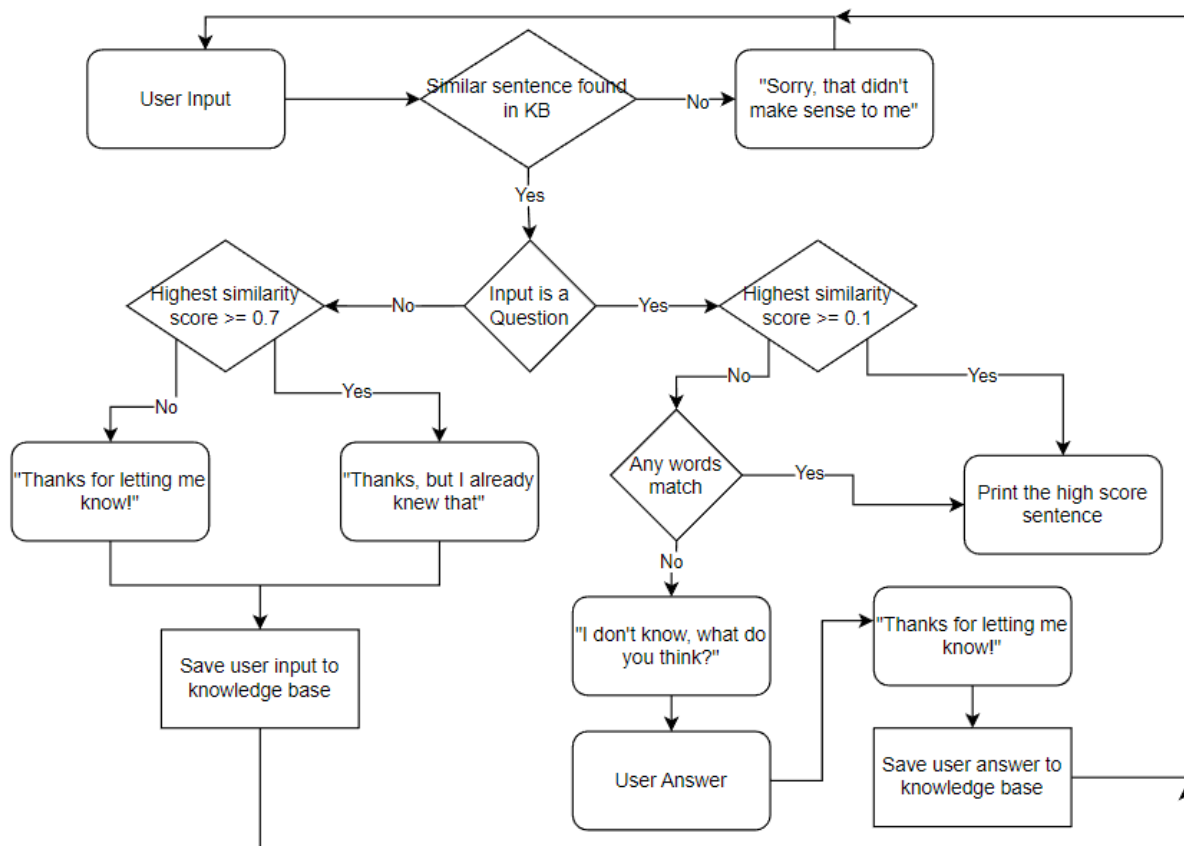
The BBall Chatbot can hold an interesting conversation about basketball with the user, enabled by having knowledge from various internet sources and tracking user input. The first part of the bot's operation is to build a topical knowledge base through web crawling and scraping. Python BeautifulSoup is used to crawl and scrape 150 relevant links starting from an initial link, excluding links that would not be pertinent, and then connecting to the others. Relevant text from each page is saved to a file, where it can then be processed. A directory of the raw text is used to create a directory of processed text. NLTK's sentence tokenizer is used to split the sentences in the text file into separate sentence tokens, and the tokens are cleaned to be useable for further analysis. All the sentences are then compiled to be a single body, and tf-idf is calculated to find the most important words, excluding stopwords. NLTK's word tokenizer is utilized to split the sentences into individual words, so term frequency can be calculated first, followed by inverse document frequency. A knowledge base of all the sentences containing words with the top tf-idf scores (the most important words) can now be created. This is done by creating a dictionary with the most important words as the key and all the sentences as values. This dictionary is then turned into a CSV for future access.

Once this CSV knowledge base is created, the chatbot reads it into a Python pandas dataframe, and the user is asked for their username. If the username does not exist, a user model is instantiated. Every time the user provides input, that input is word tokenized and checked if it is a question. Then the knowledge base data frame is searched for a sentence with high similarity. This is done by first calculating the BLEU scores of sentences in the knowledge base through NLTK, which finds the level of n-gram overlap for two sentences, up to 4-gram. The user input and each sentence is compared to generate a score from 0.0 to 1.0, and a smoothing function is used as well to handle short sentences. The sentences with a score greater than 0.1, and therefore some level of similarity, are collected in a list. A deep learning Sentence Transformer model is now loaded and the list of candidate sentences are provided to it. The model is pretrained to calculate the sentence similarity between the input sentence and the candidate sentences, and the best match sentence is returned. If no match is available, the bot will respond that it doesn't understand and ask for new input.

The chatbot now responds based on whether or not a question was asked and how high the similarity score is. If the user asked a question and the similarity score of the most similar sentence found is more than 0.1, the bot will respond with that sentence. If the score is less than that, the input will be compared with the similar sentence for any matching words. If there is a match, the bot will print the similar sentence. If not, bot will say it doesn't know and ask the user for their opinion. This response will be saved in the user model for future reference. On the other hand, if the user input was not a question, the bot will still look for the most similar sentence and then respond that it already knew if the similarity score is more than 0.6, or thank the user for the new information if the score was less than that. The new information will also

be saved to the user model. The sentences added to the user model will also be checked for similarity in the future, so the bot will show that it remembers what the user has said to it. The user models are part of the Chatbot object, which is pickled to maintain persistent user models that can be accessed with a username.

Diagram of Dialog Tree



Sample Dialog Interactions

Sample 1

>> Hi! What is your name?

You: Kevin

{'Kevin': <__main__.UserModel object at 0x000001F2189AEB50>}

>> Welcome back, Kevin! This is a chatbot for all things basketball related!

>> Type 'stop' to quit.

You:Who is the greatest basketball player ever?

>> 1961 1966 1967 1968 1970 1974 1980 1988 1989 1995 2004

Vincent Lamar Carter Jr. (born January 26, 1977) is an American former professional basketball player who serves as a basketball analyst for ESPN.

You:Is Michael Jordan the greatest basketball player ever?

>> Michael Jordan is the greatest basketball player ever.

You:Who is the greatest basketball player of all time?

>> 1961 1966 1967 1968 1970 1974 1980 1988 1989 1995 2004

Vincent Lamar Carter Jr. (born January 26, 1977) is an American former professional basketball player who serves as a basketball analyst for ESPN.

You:That is wrong.

>> Sorry, that didn't really make sense to me.

You:That is wrong. Michael Jordan is the greatest basketball player of all time.

>>Thanks, but I already knew that. Michael Jordan is the greatest basketball player ever.

You:stop

Process finished with exit code 0

Sample 2

>> Hi! What is your name?

You: Kevin

```
{'Kevin': <__main__.UserModel object at 0x0000027D30F6DB50>}
```

>> Welcome back, Kevin! This is a chatbot for all things basketball related!

>> Type 'stop' to quit.

You:What are some basketball moves?

>> Basketball courts have a three-point arc at both baskets.

You:What are some techniques of basketball?

>> Carrying is a violation in the game of basketball.

You:Dribbling is a part of basketball.

>>Thanks, but I already knew that. Dribbling was not part of the original game, but was introduced in 1901.

You:How do you dribble a basketball?

>> Dribbling is a part of basketball.

You:I know that, but I want to know what the action of dribbling is.

>> Sorry, that didn't really make sense to me.

You:I don't think you actually know how to dribble.

>>Wow! Thanks for letting me know!

You:

Appendix A: Knowledge Base

The knowledge base was generated by sentence tokenizing and processing the web scraped text. A dictionary of the most important words, determined by tf-idf, was created, and the sentences containing the keywords were added as values. This was then converted to a CSV for access through a data frame.

Here is a sample of the CSV that serves as the knowledge base:

dribble,"In starting a dribble after (1) receiving the ball while standing still, or (2) coming to a legal stop, the ball must be out of the player's hand before the pivot foot is raised off the floor."

dribble,"The Euro-Step itself involves picking up one's dribble while dribbling, taking one step in one direction, then quickly taking a step in the other direction to avoid the defender to create room for a layup attempt."

dribble,"A skilled player can dribble without watching the ball, using the dribbling motion or peripheral vision to keep track of the ball's location."

dribble,"A dribble that touches both hands is considered stopping the dribble, giving this infraction the name double dribble."

carter,Carter was named an Eastern Conference All-Star in 2006.

carter,"For the latter half of April and the whole first round playoff series against the Spurs, Carter was inserted in the starting lineup and played well."

carter,Carter would finish with 39 points in a 101–90 New Jersey victory.

carter,"On March 13, Carter made his first start of the season and made all eight of his shots, including six from beyond the arc, to score a season-high 24 points and lead the Grizzlies past the Milwaukee Bucks 113–93."

carter,"At 40 years, 46 days old, Carter also became the oldest player to start an NBA game since Juwan Howard in April 2013."

guard,"The point guard is positioned on the perimeter of the play, so as to have the best view of the action."

guard,"Lee Rose has described a point guard as a coach on the floor, who can handle and distribute the ball to teammates."

guard,"If the shooting guard's main priority is to limit or prevent the opposing team's star player (which is usually another shooting guard or other perimeter player), then the shooting guard could be known as a defensive specialist."

guard,A point guard that has enough ball-handling skill and quickness to be able to drive to the basket in a half-court set is also very valuable and considered by some to be a must for a successful offense.

guard,A point guard should always look out for teammates and be able to pass well.

Appendix B: User Models

The user models are maintained within the Chatbot object, usernames are also here:

```
{'Kevin': <main.UserModel object at 0x000001F5E086EB50>, 'Hrithik': <main.UserModel object  
at 0x000001F5E086E1C0>}
```

They contain user input the system has decided to keep for future reference, since a good match was not available within the existing knowledge base. These sentences are also referenced as part of the knowledge base.

Example:

```
Dribbling is a part of basketball.
```

```
I don't think you actually know how to dribble.
```

```
That is wrong. Michael Jordan is the greatest basketball player of all time.
```

Chatbot Evaluations and Analysis

One of the strengths of this chatbot is that the text it produces comes from simple information retrieval. Because of this, the output of the chatbot is almost always grammatically correct. Furthermore, when there is a good match between the input sentence and a sentence in the knowledge base, the output of the chatbot can approach a human likeness because it is selecting from sentences literally generated by humans, for humans. Furthermore, included within is a pretrained model using a deep neural network (BERT, a transformer model) capable of producing similarity scores between two sequences. Although this score takes a relatively long time to compute, the chatbot does not take long to respond because we also use the BLEU

metric as a heuristic to collect sequences from the knowledge base that are generally similar to the user input sequence, filtering out sequences with little to no relevance to what the user is saying. Furthermore, generative models require more sophisticated datasets; for example, the training process requires that the generative model knows what the output is supposed to look like, meaning that we need specific datasets with the correct outputs. A sufficiently sophisticated neural network trained on a sufficiently large dataset would also be large and take a long time to train, which is a problem our knowledge retrieval bot avoids. Because it has a minimal set of rules, our chatbot is also easier to code and does not require a lengthy brainstorming phase where rules need to be generated and ordered correctly.

On the other hand, when it does not find a match the model is forced to either settle for a tangentially related sequence in the knowledge base to output, or if there are no related sequences, the model has to output a formulaic set of sequences, asking the user for more user input to add to the knowledge base. To resolve this issue, it might make more sense to add even more sequences to our knowledge base, but at a certain point the size of the knowledge base will inevitably lead to problems with runtimes. This means that we can't simply increase the size of the knowledge base ad infinitum and expect the model to be better, because users will expect the chatbot to respond within a timeframe. Furthermore, because we select the most similar sentence, with no extra output the chatbot will return the exact same sentence each time. We could add a level of randomness by selecting the top-k most similar sentences, but this would require a large and varied enough dataset such that the top-k sentences are all coherent and relevant, which was not the case for our knowledge base.