ChatRegex for Detective Novels – COSC 524 Group 6

Matthew Puryear, Karen Hughes, Brandan Roachell, Drew Friend October 15, 2023

1 Introduction

For this project we used Natural Language Processing to analyze the detective novels *The Hound of the Baskervilles* [1] by Arthur Conan Doyle as well as *The Mysterious Affair at Styles* [2] and *The Murder on the Links* [3], both by Agatha Christie. The goal of the project was to develop a chatbot that was able to answer questions about the novels' plots, settings, characters, and the relationships thereof. To do this, the program had to be able to parse both the novels and the questions asked of it and then assemble natural-sounding responses to accurately answer those questions.

2 Design Choices

Our first major design choice was generalizing the prompt questions. To start, all six of the required prompt questions can be broken down into thee main categories. These categories are when a person/category classification first occurs, the words surrounding a person/category classification, and the co-occurrence of people/category classifications. These three question types are then determined through a variety of keyword searches to determine the final question response type. From here, prompt responses were selected based on how people and/or category classifications are requested. More specifically, if the question asks about a general category only, then the response individually covers everyone in that category. Likewise, questions about a single person will only have responses regarding that person. Most characters in these books only go by one name. If they go by two names, both names would map to the person when searching the prompt so the result would be the same for both names. If the prompt mentions a category and a person in that category (for example, "detective Holmes"), the interpreter treats this as a single person for questions pertaining to single person answers but only compares the name to other names for co-occurrence prompts within their category classification. In essence, co-occurrence analysis can be performed between two categories, two people, or a person and a category even if they are part of the category in question. By designing the prompt interpreter in this way, it could be made to cover all the required questions while being able to handle a much broader scope of question subjects.

Before beginning to parse the texts, they were manually pre-processed to remove unimportant information. This is because Project Gutenberg includes a lengthy header and footer on each text that covers the legal needs of the service provided. Each raw text that was used from Project Gutenberg was stripped of all sections that weren't part of the story itself and added a new file for use. This pre-processing step also allowed us to standardize the way that chapters were delineated between works to making it easier to generalize the analysis methods across them.

One of the first steps when parsing the text was to break it up into chapters of sentences. To start, chapters and paragraphs were picked out based on spacing. Then, sentences were divided within paragraphs based on the presence of punctuation. Unfortunately, looking for punctuation alone is too simple and naive for this task. When splitting on each punctuation mark that could denote the end of a sentence, titles like "Dr." would affect the parsing. To counteract this, we made a regex pattern to identify true sentence ending punctuation ensure that it is not part of a title using negative look-behinds. From here, the paragraph-sentence structure was flattened to produce a single list of sentences for each chapter with the first sentence being the chapter title.

Once the books are successfully split into sentences, each sentence is processed to identify character and crime match locations. This is done using regular expressions individually crafted to uniquely identify the major crime and each major character in the following categories: detectives, perpetrators, victims, and suspects. Occasionally, there are sentences where the identity of the object or subject is unclear when based solely on the given name. That could be as a result of missing honorifics/titles, lack of context from outside the sentence, or last names shared by multiple characters. Because of this, it was decided that in these cases would not attributed to any characters. This was determined to be the lesser of two evils when compared to possibly identifying these names with wrong character.

3 Findings

The frequency of detectives, perpetrators, and suspects were analyzed to determine interesting relationships between the characters and the plot.

The Hound of the Baskervilles consists of two detectives. The detective Sherlock Holmes and Dr. John Watson occur 185 and 114 times, respectively. The perpetrator Jack Stapleton occurs 68 times. The other top three suspects Henry Baskerville, James Mortimer, and the Barrymores were mentioned 145, 90, and 79 times, respectively.

For *The Murder on the Links*, there are five detectives. Hercule Poirot, Arthur Hastings, Monsieur Giraud, Monsieur Hautet, and Japp were mentioned 470, 63, 161, 70, and 6 times, respectively. Regarding the perpetrator, Marthe Daubreuil was mentioned 68 times. This is significantly less often than the two other main suspects, Jeanne Beroldy and Jack Renauld, who were mentioned 112 and 157 times, respectively.

In *The Mysterious Affair at Styles*, there are three detectives Hercule Poirot, Arthur Hastings, and Japp with an occurrence frequency of 380, 48, and 25, respectively. The perpetrators Alfred Inglethorp and Evelyn Howard occur 67 and 20 times, respectively. Most of the other suspects

occurred more times ranging from 57 to 173 times for the five suspects.

Overall throughout all three novels (across two authors), it seems that other suspects appear significantly more often than the true perpetrator(s) by design—perhaps to allow them to slip from the mind of the reader by having less focus.

When comparing between when the first time the main detective, perpetrator(s), crime, and suspects were introduced, we noticed that the detective was always introduced at the beginning of chapter 1 for all three books while the perpetrator and suspects are first mentioned sometime after, typically in a later chapter entirely. In both of the Agatha Christie novels, the crime was first mentioned in chapter 3. It would be interesting to analyze additional books by this author to see if they follow this trend. In contrast, Doyle doesn't reveal the details of the crime itself until chapter 15 in *The Hound of the Baskervilles*. It is unclear whether he consistently keeps the crime a mystery in his works.

As for the suspects and perpetrator(s), *The Mysterious Affair at Styles* first mentions the perpetrators in chapter 1—before the crime—along with the other suspects while *The Murder on the Links* doesn't mention the perpetrator until chapter 7 but does introduce the other suspects in close proximity to the crime (chapters 3 and 4). While Christie does not have a consistent style for when the perpetrator is introduced, it is interesting how they appear to either be "hidden" among the other suspects or distanced from the crime. In *The Hound of the Baskervilles*, the perpetrator and other suspects are mostly mentioned around chapters 1 and 2 with a few mentioned several chapters later.

We also analyzed the most common words surrounding many of the characters, but nothing seemed significant enough to be worth noting, and no words stood out.

4 Challenges

Titles caused numerous issues for us. The first was intentionally leaving in the periods following titles such as "Mr." and "Mrs." while parsing into sentences. This was resolved as mentioned in the Design Choices section above. The issues brought about by titles continued though, as there were sometimes characters with the same last name—notably family members. When titles or honorifics were used instead of first names, it became unclear at the parsing level who was being mentioned. Depending on the makeup of the family, titles can occasionally help differentiate, but they are still less informative than using full names. This is, of course, necessary to make realistic dialogue and human narration, and being able to derive who a name or pronoun is referring to from context is a challenge for more advanced natural language processing techniques to accomplish. In our case, we simply erred on the side of caution in our analysis and disregarded any instances of ambiguous character names unless they were the only one referred to as such, which we confirmed by manually reading the usage of that name throughout each novel. The full list of accepted ways to refer to a character can be found in `analyze_books.py`.

We also faced challenges when dealing with quotations and dialogue. Punctuation and sentence structure are notably altered when either of these are present. Our current implementation, in order to stick most accurately to the correct number of sentences, is to have the quote (or first sentence of the quote if it is separated into multiple parts) be considered as part of the sentence fragment before it, and to have any sentence fragment when the quote ends be marked as a unique sentence. In future work, we had an idea about how to fix this with a hierarchy of sentence structure. When seeing quotation marks, the program would note that it was in a "secondary" sentence. Then, when the quotation is finished, it can return to the "primary" sentence that was not previously exited. This would allow reporting of quotes within a sentence as "sentence 10.2", for example, for the second sentence within a piece of dialogue contained within the tenth sentence of the chapter.

For the analysis of *The Hound of the Baskervilles*, there were challenges in determining when the crime was first mentioned due to the type of crime and its details not being fully revealed to the readers. This novel involves a superstitious component with legend of a hound plaguing the family. It it revealed that the victim indeed died from a heart condition as earlier mentioned from the medical examination.

The most ongoing challenge that we faced when dealing with prompts was handling synonyms. In order to naturally answer questions that could be worded in many different ways, we created large patterns of synonyms for the keywords within questions, such as "occur," "mentioned," "seen," etc. This came with issues, however, when words could feasibly be used to ask very different questions. As long as the user is asking questions concisely and clearly, not attempting to trick the program by hitting keywords picked up by multiple questions (e.g., the prompt "Around when does the detective first appear?" would try to answer about surrounding words as well as when the character was first introduced), our implementation is quite successful, but we could handle prompts more flexibly given additional time.

5 Conclusion

Our project was successful in its goal of being able to answer questions about the assigned books beyond that which was required. Using the "anchor" of specific characters and events, the program can effectively parse through the story to respond to specific queries about the narrative.

In general, it does not seem like analyzing the plot structure alone can predict the final outcome of a novel. While the program does provide interesting details, unless a perpetrator is explicitly revealed in context, the results do not feel conclusive enough even with deeper analysis. Our most noteworthy finding is how infrequently a perpetrator is mentioned relative to other main suspects and may be a trend worth investigating other detective novels for.

6 References

[1] A. C. Doyle, *The Hound of the Baskervilles & the Valley of Fear*. Pan Macmillan, 2016, vol. 24.

[2] A. Christie, *The Mysterious Affair at Styles*. e-artnow, 2018.

[3] ——, *The Murder on the Links: A Hercule Poirot Novel*. Berkley, 1923, vol. 2.