FRI Natural language processing course 2022
Final report

University *of Ljubljana*
Faculty *of Computer and*
Information Science

Literacy situation models knowledge base creation

Andrej Drofenik, Enej Bačić and Sebastjan Tkavc

Abstract

In natural language processing, the task of reliably analyzing or interpreting literary works can be very challenging for machines, despite tremendous developments in the field. These challenges arise from the machines' inability to understand and interpret contextual information the way humans do, lacking the appropriate knowledge and background information. In our project, we focused on character analysis, which is a popular task in the field of literary analysis. Specifically, we tackled the problem of extracting characters from short stories and determining their respective protagonists and antagonists. To achieve this, we used a number of datasets and techniques of analysis.

Keywords

character extraction, named entity recognition, sentiment analysis, machine learning, short stories ...

Advisors: Slavko Žitnik

Introduction

In the field of natural language processing, researchers have tackled many different problems to help machines understand written language. Since literature presents an easy method of obtaining suitable datasets for analysis, many projects have been developed to tackle problems that are specific to literary works. During the very long history of literature across the world, all works have been written for human understanding, which is very different from how machines interpret written language. While a person can read a story and instantly recognize the story's themes, main characters, locations where the story takes place and sequences of important events, a machine has to be taught to do so in a completely different manner.

In this work, we attempted to discover the main characters in groups of short stories and determine the protagonists and antagonists through a series of NLP-based approaches. For this purpose, we used two different datasets, containing a number of short stories of varying lengths, vocabularic complexities and writing styles. For the stories in the datasets, we used named entity recognition (NER) to extract the characters' names, which we combined with sentiment analysis using the BERT language model [1]. With this approach, we were able to determine the protagonists and antagonists of short stories with varying levels of success.

Related work

Extensive research has been done on narrative comprehension, the requirements of situation models and implementation of such models.

A recent study [2] has shown that local semantic relations significantly influence recall of paired sentences in L2 readers. Results show that the global casual relations and local semantic relations have a large impact on a reader's memory. This provides insight to assessing text meaning.

As semantic relations are beneficial to meaning assessment in text, we looked at an overview of the field of semantic analysis in natural language processing. The paper by Salloum et al. [3] provides an insight on methods such as latent semantic analysis, explicit semantic analysis and sentiment analysis and the overall importance of semantic analysis.

The paper by Zwaan et al. [4] researched the importance of three dimensions of situational continuity. These consist of temporal, spatial and casual continuity. The authors have also shown that readers simultaneously monitor more than one dimension under normal reading instruction, temporal and casual having the most impact.

In the work of Dasgupta et al. [5] automatic extraction of cause-effect relations using their proposed bi-directional LSTM model with an additional linguistic layer. They achieved better performance than other methods. A product of the research is also an annotated dataset in the sense of cause-effect relations.

The paper Extraction and Analysis of Fictional Character

Networks: A Survey [6] provides information on the entire process of character networks. It explains the steps necessary to construct such a network, such as character identification, interaction detection, graph extraction. It outlines the current situation in the field, the methods and performance.

In the work Exploring Cross-sentence Contexts for Named Entity Recognition with BERT [7] the authors describe the importance of context, that is usually found in different sentences that the named entity and perform test and evaluation of prediction of the sentences in different contexts.

Datasets

Our analysis was based on two datasets of short stories. We were initially provided with the IMapBook dataset, which includes 7 short stories in plain text format:

- ”The Ransom of Red Chief” by O. Henry (4162 words),
- ”Hills Like White Elephants” by E. Hemingway (1473 words),
- ”Leiningen vs. the Ants” by C. Stephenson (8663 words),
- ”The Lady or the Tiger” by F. R. Stockton (2701 words),
- ”The Most Dangerous Game” by R. Connell (8009 words),
- ”The Tell-Tale Heart” by E. A. Poe (2209 words),
- ”The Gift of the Magi” by O. Henry (1865 words).

The listed short stories do not share any common characteristics that would make it unsuitable for our analysis. In fact, they vary immensely in narrative technique, writing style and most importantly, the selection of characters portrayed within. ”The Most Dangerous Game” has many characters which are directly referenced by their name several times, while ”Leiningen vs. the Ants” features only one. ”The Lady or the Tiger” and ”The Tell-Tale Heart” feature no named characters, which makes any analysis concerning characters quite difficult. Due to the short length of these stories, any sentiment analysis would also be more unreliable if the characters in the stories were referenced a small number of times. Therefore, it would be more difficult to determine the protagonists and antagonists correctly.

We chose our second dataset in a way to mitigate the issues presented in the IMapBook dataset. For this purpose, we selected 12 short stories which were freely available from Project Gutenberg [8]. These include:

- three detective novels by Sir Arthur Conan Doyle, featuring the famous Sherlock Holmes character,
- three fantasy novels by Robert E. Howard, featuring Conan the Barbarian,
- three stories by Rudyard Kipling, including ”The Jungle Book”,
- three horror stories by H. P. Lovecraft, including ”The Call of Cthulhu”.

These stories are substantially longer, ranging from approximately 7600 words to approximately 60000. The selected texts also differ quite greatly in writing style, with Kipling’s stories being more children oriented, while Lovecraft’s writing is much more indirect and uses a more expansive vocabulary. Unlike the IMapBook dataset, the stories from Project Gutenberg would provide us with a more reliable testing ground for our methods of analysis.

Methods

This section describes the methods we used during our work. As described, our goal was to determine the protagonists and antagonists in each short story. To begin, we stripped the short stories of their chapter markings and other text that was not part of the story, such as the forewords.

Named Entity Recognition

Since our plan was to determine what characters appear in these stories, a good starting point was the NER (Named Entity Recognition) technique. With this approach, we were be able to determine all named entities in a given short story. These also include entities that do not represent characters, which were of little relevance for our goals.

We implemented the NER technique with the spaCy framework [9], an open source NLP software. The NER model included in the framework uses a small English pipeline trained on written web text (blogs, news, comments), that includes vocabulary, syntax and entities. This technique begins by tokenizing the input text and returning a prediction for each token that could represent an entity. By using the spaCy framework, we were able to reduce the predicted list of entities to only those that represent persons, which resulted in a comprehensive list of characters for each short story. During analysis of longer texts, this list is reduced further to include only characters with unique names. We also counted the number of appearances for each character, which would provide a simple way of determining which characters are often referenced in the stories.

From a reliability perspective, the reliance on the framework to reliably reduce the list only to persons is very questionable. As we described in the datasets section, certain short stories from the IMapBook dataset do not include any characters that are referred to by their first or last name. Even for stories that include many named characters, this approach can completely disregard any character if it is not referenced by a unique name. We predicted that this simple NER implementation would produce better results on the more expansive short stories, such as the ones from the Gutenberg dataset.

Sentiment analysis on the results from NER

For the sentiment analysis we have saved the 4 most frequent named entities from NER for every short story in our two datasets. With this data we have included the immediate appearances of these entities (the sentences that contain them) as well as their extended context. We have included 5 sentences preceding the appearance and 5 sentences following it. This allowed us to perform analysis with a variety of context options.

The following options were used and evaluated during our analysis:

- immediate appearance (only the sentence where the named entity is detected),
- immediate appearance and the following 3 sentences,
- immediate appearance and the following 5 sentences,
- immediate appearance and the preceding 3 sentences,
- immediate appearance and the preceding 5 sentences,
- preceding 3 sentences, the immediate appearance and the following 5 sentences.

The model used for sentiment analysis was a fine-tuned checkpoint of RoBERTa-large, that was fine-tuned and evaluated on different types of texts [10].

We have calculated the average negative and positive scores for all appearances over the above context options.

For protagonist and antagonist detection we have made two large assumptions:

- the main characters are the ones most frequent in the short story and
- the average context of the character is enough to determine their status.

For result interpretation we visualized the data for every detected character in the stories from our datasets.

Diagram of analysis

In Figure 1 we show the outline of the data manipulation and analysis process. As mentioned above we run named entity recognition on the dataset stories. Those entities are then equipped with their context (immediate appearance and 5 sentences in each direction) and used in sentiment analysis. When the sentiment analysis is performed a detection of protagonist and antagonist is evaluated using some assumptions, sentiment and frequency data. The results are then visualized for easier interpretation. At every step the results are saved to disk, to enable further comparison and analysis of the already computed data.

Results

In this section, we present the results of our work.

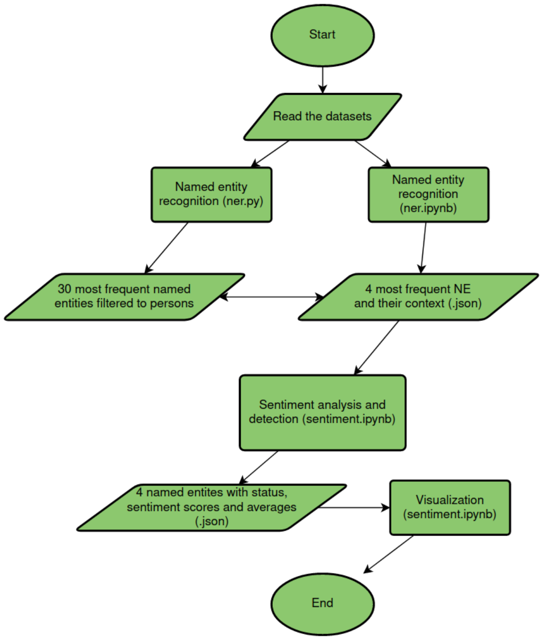
Start
Read the datasets
Named entity recognition (ner.py)
Named entity recognition (ner.ipynb)
30 most frequent named entities filtered to persons
4 most frequent NE and their context (.json)
Sentiment analysis and detection (sentiment.ipynb)
4 named entites with status, sentiment scores and averages (.json)
Visualization (sentiment.ipynb)
End

Figure 1. Diagram showing the outline of the knowledge base creation.

Named entity recognition

In table 1, we provide the results of using this technique on Conan Doyle's ”The Sign of the Four”, approximately 44.000 words in length.

Table 1. Named entity recognition

Entity Name	No. of occurrences
holmes	134
sholto	76
morstan	70
jones	44
thaddeus	36
sherlock	33
smith	27
toby	26
watson	24
bartholomew	22

The results, while flawed, can be very useful when attempting to determine the main characters in a story. By observing the number of occurrences, we can see a clear difference between the main characters of a story and side characters that the story isn't centered around. Towards the bottom of the list, characters are mentioned at a much lower rate than the top, suggesting that they play a lesser role in the story.

While this simple analysis can be useful to an extent, the implementation is flawed and requires manual oversight to function properly. An easily observable issue is the inclusion of both 'holmes' and 'sherlock' values in table 1 as separate entities, even though they are referencing the same person.

Stories where characters are referenced by their name and surname separately are therefore problematic, as there is no simple way of determining which character is being referenced based solely on either one. Two characters usually do not share the same first name, but characters sharing surnames is quite common (in this case, Sherlock and Mycroft Holmes). Ultimately, the method was not changed, since we determined that future results when determining protagonists and antagonists would not be affected by this flaw.

The IMapBook dataset proved to be a much tougher challenge for the NER technique. The issues we outlined in the datasets section impacted our results significantly, since the implemented NER technique could not effectively process three of the four short stories. ”Hills Like White Elephants”, ”The Lady or the Tiger” and ”The Tell-Tale Heart” included no named entities which would reference characters in the story, therefore it produced empty lists. In the same light, ”Leiningen vs. the Ants” has only one true character, which the method successfully recognizes, but fails to detect anything else. This NER implementation therefore is not suitable for any character extraction when the characters are not explicitly referenced by their names. We recognize this as a significant flaw in our framework, since further analysis for these stories is not possible due to the implementation’s shortcomings.

Sentiment analysis

Before we could obtain sentiment analysis and protagonist and antagonist detection we needed to evaluate the best context option to use. We have run the sentiment analysis over all our examples for every defined option. With the results we could observe that the following 5 sentences including the immediate appearance provided the best average scores overall as seen in Figure 2. But our decision was to use the most complex context option as it did not fall far behind the leading score but proved somewhat more stable in the final results.

Using our decided context option we have labeled the sentiment of all the detected named entities across the stories.

The result for the most frequent NE that were included in the top 4 most frequent NE of a story can be seen in Table 2.

Table 2. Named entity recognition - sentiment analysis results

Entity Name	#	Sentiment
holmes	134	POSITIVE - Protagonist
sholto	76	NEGATIVE
morstan	70	POSITIVE
jones	44	N/A
thaddeus	36	N/A
sherlock	33	N/A
smith	27	N/A
toby	26	N/A
watson	24	N/A
bartholomew	22	N/A

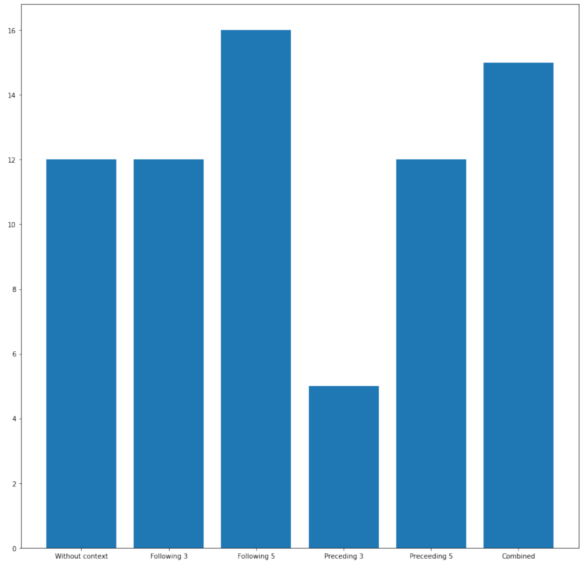

Figure 2. Bar graph analysis of using different context options with sentiment analysis.

From the above example we can see that the protagonist was properly detected. But in the results we can also observe that a non person named entity was detected and wrongly classified as seen in Figure 3. Even though the relation was not properly detected, we can still observe that the main character was properly recognized.

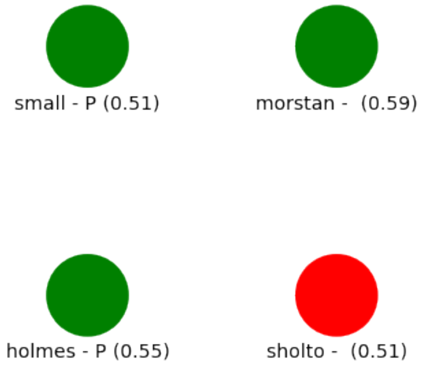

Figure 3. Analysis of the characters in the story ”The Sign of the Four” by Conan Doyle.

A more successful detection can be seen for the story ”SHADOWS IN ZAMBOULA” by Robert E. Howard. Here the characters were identified and frequent enough for our assumptions to hold as seen in Figure 4

As mentioned in the previous section the detection on the IMapBook dataset was a much more difficult problem with the main characters being hard to detect and also having very vague context spread out across almost the entire short story.

More random text

Discussion

We are somewhat satisfied with the results we achieved during this project. By using simple NLP methods we were able

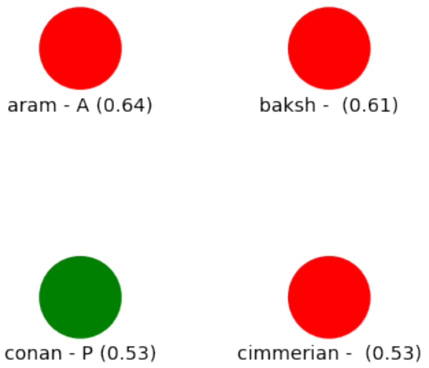

Figure 4. Analysis of the characters in the story ”SHADOWS IN ZAMBOULA” by Robert E. Howard.

to extract the characters from appropriate short stories and determine the protagonist and antagonist with reasonable accuracy. There are some advantages to the implementations and frameworks we chose to explore, however many of the shortcomings of our work could be remedied by the use of additional frameworks and techniques that would complement our current implementation.

Starting with NER, our implementation using the spaCy framework is only suitable for stories where characters are explicitly referred to by their first or last names. Since some of the stories from our datasets lacked named characters, we were unable to extract any of them in an automatic manner. This is by far the biggest flaw in our work, since it prevents any further analysis for these short stories. Furthermore, when detecting named characters that are referred to by their first and last name at different points, it can be very difficult to separate characters that share either of those two names. This would be the first and easiest improvement we would make to our framework, since it would make the NER perform very well for named characters.

Our method of protagonist and antagonist detection is flawed as it do not take into account the relations between the characters. This could be much better handled with a deep learning approach instead of hand crafted decisions. Our preferred method would be to use the RoBERTa model and fine-tune it to learn the classification into protagonist or antagonist based on the sentiment of the character in it’s context. To achieve this we would need a larger dataset, with more extensive annotations.

We could also improve the detection and sentiment analysis by linking characters within the stories and taking their relations into account. This could be achieved with using self-attention of a BERT model and creating a custom classification layer that would return the highest attention tokens based on the inputted character. But we ran into the problem of even these short stories being too long to use with a BERT model. We have explored the possibility of using something like Longformer, a transformer that can handle long documents [11].

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Shingo Nahatame. Revisiting second language readers’ memory for narrative texts: The role of causal and semantic text relations. *Reading Psychology*, 41(8):753–777, 2020.

[3] Said A. Salloum, Rehan Khan, and Khaled Shaalan. A survey of semantic analysis approaches. In Aboul-Ella Hassanien, Ahmad Taher Azar, Tarek Gaber, Diego Oliva, and Fahmy M. Tolba, editors, *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, pages 61–70, Cham, 2020. Springer International Publishing.

[4] Rolf A Zwaan, Joseph P Magliano, and Arthur C Graesser. Dimensions of situation model construction in narrative comprehension. *Journal of experimental psychology: Learning, memory, and cognition*, 21(2):386, 1995.

[5] Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 306–316, 2018.

[6] Vincent Labatut and Xavier Bost. Extraction and analysis of fictional character networks. *ACM Computing Surveys*, 52(5):1–40, Sep 2020.

[7] Jouni Luoma and Sampo Pyysalo. Exploring cross-sentence contexts for named entity recognition with BERT. *CoRR*, abs/2006.01563, 2020.

[8] Project gutenberg.

[9] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[10] siebert/sentiment-roberta-large-english - hugging face.

[11] allenai/longformer-base-4096 - hugging face.