FRI Natural language processing course 2021
Project report

University of Ljubljana
Faculty of Computer and Information Science

Challenges in Creating a Knowledge Base for Literacy Situations: Character Extraction and Analysis

Matic Šuc and Loris Štrosar Grmek

Abstract

Keywords
named entity recognition, sentiment analysis

Advisors: Slavko Žitnik

Introduction

Natural Language Processing (NLP) is a subfield of Artificial Intelligence (AI) that focuses on enabling computers to understand and generate human language. One of the main challenges in NLP is to build systems that can reason about text in a way that mimics human cognitive abilities. To achieve this goal, researchers have developed a variety of models that capture different aspects of language understanding, such as syntax, semantics, and pragmatics. One such model is the Literacy Situation Model (LSM), which captures the knowledge and assumptions that people use to understand written text. The LSM is based on the idea that readers construct mental models of the situations described in texts, and that these models help them to interpret and remember the information conveyed. Creating a knowledge base for LSMs involves extracting and organizing the relevant information from texts, such as entities, events, and relationships, and encoding it in a structured format that can be used for reasoning and inference. This process involves a range of NLP techniques, such as named entity recognition, coreference resolution, and semantic role labeling. The resulting knowledge base can be used for a variety of applications, such as text understanding, question answering, and summarization.

Related work

Corpus Construction

In [1], the authors introduce the ROCStories corpus, consisting of 50,000 commonsense stories. This corpus is particularly notable as it serves as the foundation for the Story Cloze Test evaluation framework, enabling deeper understanding of commonsense stories.

The annotated dataset of literary entities presented in [2] offers valuable resources for character analysis and entity recognition. Derived from 100 English literary texts, this corpus provides comprehensive annotations for various entity categories, facilitating research on entity recognition and character-centric NLP tasks. Additionally, the ACE entity annotated dataset, discussed in the same paper, serves as a significant contribution to named entity recognition (NER), offering extensive annotations for named entities across multiple categories. This benchmark dataset not only enables the evaluation of NER models but also advances research in this field, driving the development of more accurate and robust NER systems.

Moreover, in the realm of corpus construction, the paper [3] presents a hybrid approach to relationship extraction from stories. By combining unsupervised and supervised learning methods, the authors propose a methodology for identifying main characters and extracting their relationships. The study utilizes a collection of 100 short stories for analysis, resulting in the identification and categorization of 300 character pair relations. This approach proves to be effective in capturing various types of relationships, such as 'Parent-Child' or 'Friendship', showcasing its significance in corpus construction for character analysis and relationship extraction.

NLP Task Advancements

In the context of NLP task advancements, already mentioned research paper [2] presents a hybrid approach to relationship extraction from stories. This methodology effectively identifies the main characters in a story and extracts their relationships, offering valuable insights for character analysis. By leveraging a combination of unsupervised and supervised learning methods, this approach demonstrates robust performance in accurately identifying and capturing character relationships within narrative texts.

The work by Dasgupta et al. [4] focuses on automatic extraction of causal relations from text using linguistically informed deep neural networks. By incorporating linguistic features and employing deep learning models, this approach enhances the understanding of causal relationships between entities, advancing the field of NLP.

The authors in [5] address the task of summarizing short stories. Their approach involves extracting important entities and events from the stories to create summaries that provide users with relevant information about the story's setting without revealing the plot. This research contributes to the development of effective techniques for summarization and aids users in decision-making regarding reading choices.

In [6], the CoNLL 2003 shared task is introduced, providing benchmark datasets for multilingual named entity recognition (NER). This work plays a crucial role in advancing NER research across different languages by offering standardized evaluation datasets and promoting cross-lingual NER studies.

The Stanford Sentiment Treebank, presented in [7], offers a large-scale corpus with fine-grained sentiment annotations. This corpus facilitates research in sentiment analysis by providing sentence-level and phrase-level sentiment labels, enabling more nuanced sentiment analysis and the study of sentiment compositionality.

The work by Ma et al. [8] introduces character-aware neural language models and presents the Persona-Chat dataset. This dataset consists of conversational dialogues with annotated speaker information, enabling character analysis in a dialogue context and contributing to advancements in conversational NLP tasks.

By leveraging the aforementioned corpora and advancements in NLP tasks, researchers have made significant progress in fields such as named entity recognition, sentiment analysis, and character analysis, enabling deeper understanding of textual data and fostering the development of more robust NLP models and algorithms.

Dataset

For our study, we utilized a dataset consisting of 55 fables attributed to Aesop. Aesop's fables are a collection of short stories traditionally attributed to Aesop, a Greek storyteller believed to have lived in the 6th century BCE. These fables typically feature anthropomorphic animals, mythical creatures, and occasionally humans as characters. Each fable conveys a moral or lesson, often presented through the interactions and dilemmas faced by the characters within the story.

The fables of Aesop are known for their brevity, usually ranging from a few sentences to a few paragraphs in length. Despite their concise nature, these tales effectively convey moral teachings and reflect upon human behavior and societal dynamics. The characters within the fables serve as representatives of certain traits, virtues, or vices, allowing readers to draw parallels to real-life situations.

To facilitate our analysis, all fables in our dataset were pre-annotated. However, we found that some of the pre-existing annotations did not align with our desired criteria, so we made necessary adjustments. These modifications were made based on our interpretation and understanding of the fables, ensuring the annotations better aligned with the intended protagonist/antagonist relationships and sentiments depicted in the narratives.

The annotations are stored in a structured format using JSON (JavaScript Object Notation). This allowed us to represent various aspects of the fables, including the identification of characters involved, the assignment of protagonist and antagonist roles, and the characterization of sentiments between the characters. Sentiments were categorized into three levels: -1 for negative, 0 for neutral, and 1 for positive, reflecting the emotional dynamics and interactions between the characters.

```
{
    "characters": ["fox", "cock", "dog"],
    "protagonist": "fox",
    "antagonist": "dog",
    "sentiments": {
      "fox": {
        "fox": 0,
        "cock": 1,
        "dog": -1
      },
      "cock": {
        "fox": 1,
        "cock": 0,
        "dog": 1
      },
      "dog": {
        "fox": 0,
        "cock": 0,
        "dog": 0
      }
    }
}
```

Listing 1. Annotation of *The Fox and The Cock and The Dog*in JSON format.

Listing 1 illustrates a sample annotation providing insights into the fable titled "The Fox and The Cock and The Dog." This annotation captures crucial information about the characters involved in the fable, namely the fox, cock, and dog. The annotation designates the fox as the protagonist and the dog as the antagonist within the narrative.

The annotation also encompasses the sentiments expressed between the characters. Within the fable, the fox holds a neutral sentiment towards itself, a positive sentiment towards the cock, and a negative sentiment towards the dog. Conversely, the cock holds positive sentiments towards both the fox and the dog. The dog, however, maintains a neutral sentiment towards all characters.

By structuring the annotation in this manner, encompassing character identification, protagonist-antagonist designation, and the sentiment matrix, we gain a comprehensive understanding of the character dynamics and emotional interactions present in the fable. Leveraging this annotated dataset, our objective was to uncover insights into character extraction, sentiment analysis, and the dynamics of protagonist/antagonist relationships in the context of Aesop's fables. This rich annotation allows for a detailed analysis of the narrative's thematic elements, shedding light on the intricate interplay between characters and facilitating a deeper exploration of the fable's nuances.

Methods

In this section, we describe the methods employed for character extraction, sentiment analysis of character relationships, and protagonist/antagonist detection. We begin by discussing the techniques utilized for character extraction, including Named Entity Recognition (NER) and Coreference Resolution. Next, we outline the approach employed for sentiment analysis, focusing on analyzing the emotional dynamics between characters. Finally, we present the methodology used for protagonist/antagonist detection, aiming to identify the main characters and their roles in the narrative. Through these methods, we aim to gain a deeper understanding of the literary situations and extract valuable insights from the text data.

Named-Entity Recognition (NER)

Named-Entity Recognition (NER) is a critical component of our analysis, focused on character extraction, which plays a fundamental role in identifying and extracting named entities from the text. NER enables us to delve into the intricate relationships and dynamics among the characters in literary works. The basic idea behind NER is to automatically identify and classify named entities, such as names of persons, organizations, locations, and other predefined categories, within a given text. In our approach to character extraction, a crucial step in our analysis, we utilized three models from NLTK [9], SpaCy [10], and Stanza [11] libraries. By employing these models/methods, each with its unique features and capabilities, we aimed to evaluate their effectiveness in accurately recognizing and extracting character entities, laying the foundation for further analysis and interpretation of the text.

NLTK, a widely adopted NLP library, offers a comprehensive suite of tools and resources. For named entity recognition (NER), NLTK employs machine learning algorithms, including Hidden Markov Models (HMMs) and Maximum Entropy Classifiers (MECs). These models are trained on annotated data to identify named entities in text. NLTK provides pre-trained NER models that can be further fine-tuned for specific domains, ensuring flexibility in different contexts.

SpaCy, renowned for its efficiency and accuracy, provides pre-trained models for various NLP tasks, including NER. SpaCy's NER module combines rule-based matching and statistical models. Deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are utilized to capture contextual information and enhance entity recognition. This integration of rule-based and statistical approaches contributes to SpaCy's high-performance character extraction capabilities.

Stanza, developed by the Stanford NLP Group, is a state-of-the-art NLP library. Its NER module employs advanced deep learning techniques, specifically bidirectional LSTM-CRF models. These models leverage both left and right contexts of a word to extract relevant features and enhance entity recognition accuracy. Stanza's integration of contextual information enables robust identification of named entities in text.

By employing these three libraries, we aimed to evaluate their effectiveness in character extraction tasks within the context of our literary analysis domain. Our objective was to compare their performance and determine the most suitable library/model for our specific requirements.

Coreference Resolution - CR

Coreference Resolution (CR) is a crucial component in our analysis pipeline, aimed at resolving references to named entities in the text. The primary goal of CR is to identify all the expressions (e.g., pronouns, noun phrases) that refer to the same entity and connect them together. By resolving coreferences, we can establish a coherent representation of the entities mentioned in the text, which is essential for character analysis and understanding the relationships between characters.

CR heavily depends on the output of Named-Entity Recognition (NER) as it relies on correctly identifying and classifying named entities in order to establish coreference chains. The coreference resolution process typically occurs after NER, as it operates on the identified named entities and their mentions in the text.

In our approach, we employed the AllenNLP library, which offers a powerful coreference resolution model. Specifically, we utilized the *coref-spanbert-large-2020.02.27* model provided by AllenNLP [12]. This model is based on a neural network architecture that combines contextualized embeddings from BERT [13] and a span-ranking module to perform coreference resolution.

By incorporating coreference resolution into our analysis, we aimed to enhance the accuracy and consistency of character mentions throughout the text, enabling more robust character analysis and relationship extraction.

Character Sentiment Analysis

Character sentiment analysis plays a vital role in understanding the emotional dynamics and interactions between characters in a literary work. It involves determining the sentiment expressed towards or by specific characters within the text. By analyzing the sentiment associated with character interactions, we can gain insights into the overall tone, relationships, and character development within the narrative.

In our approach, we utilized the *siebert/sentiment-roberta-large-english* model for sentiment analysis [14]. This model is based on the RoBERTa architecture [13], a state-of-the-art transformer-based model for natural language processing tasks. The sentiment model is specifically trained on English text data to classify the sentiment of a given sentence into positive, negative, or neutral categories.

To perform character sentiment analysis, we adopted a sentence-level approach. We processed each sentence in the text and utilized the sentiment model to classify the sentiment expressed in that particular sentence. The sentiment model assigns a sentiment label to each sentence, indicating whether it conveys a positive, negative, or neutral sentiment.

Based on the characters mentioned in each sentence, we inferred the sentiment between them. If two or more characters appeared in a sentence, we attributed the sentiment label of that sentence to their relationship. For example, if a sentence expressed a positive sentiment and involved two characters, we inferred a positive sentiment between those characters. By aggregating the sentiment labels across the text and character interactions, we obtained an overview of the sentiment dynamics among the characters.

By employing this approach, we aimed to capture the sentiment nuances associated with character interactions and relationships. This analysis provided valuable insights into the emotional aspects of the narrative, allowing us to explore character sentiments and their impact on the overall storyline.

Results

Model	Precision	Recall	F Measure
Stanza	0.73	0.62	0.67
Stanza + CR	0.62	0.6	0.61
NLTK	0.56	0.19	0.29
NLTK + CR	0.6	0.21	0.31
Spacy	0.77	0.49	0.60
Spacy + CR	0.8	0.57	0.66

Table 1. Comparison of NER Model performances with and without Coreference Resolution.

Model	Detected protagonists	Detected antagonists	All
SiEBERT	33	23	55

Table 2. Initial results of sentiment analysis.

Visualization

In order to gain a deeper understanding of the sentiment between characters, visualizations can be a valuable tool. When it comes to analyzing fables, which typically involve fewer characters and shorter narratives, we have devised a unique approach to visualizing character interactions within sentences. Our aim is to enhance comprehension by providing a visual representation of the sentiment expressed.

One of the challenges we encountered is that most fables contain only a single sentence where two characters appear together. Consequently, these limited instances result in visualizations that lack informative value. However, Figure 1 showcases an intriguing visualization that effectively illustrates our concept of sentiment visualization.

Taking the fable ”The Lion and the Mouse” as an example, we observe that both the lion and the mouse appear in three sentences. The visualization employs three lines of varying lengths, each corresponding to the length of an individual sentence. In the first sentence, a negative sentiment is detected between the lion and the mouse, while the remaining two sentences depict a positive sentiment. Through this visualization, we gain a clearer perspective on how the sentiment changes over the course of the fable.

Figure 1. Characters sentiment visualization. The sentiment visualization of *The Lion and the Mouse*. Three separted lines representing the red negative and green positive sentiment in three sentences with both characters appearing in them.

The next enhancement to our visualization concept is showcased in Figure 2. This example focuses on the story titled *The Cat Maiden*, where both Venus and Jupiter are mentioned in two sentences exhibiting positive sentiment. To represent this, we utilize two lines, each depicted in a different shade of green. The visualization effectively illustrates that the sentiment in the first sentence is detected with a relatively lower confidence score, whereas the sentiment in the second sentence is detected with a higher confidence score.

Figure 2. Characters sentiment visualization. The sentiment visualization of *The Cat Maiden*. Two green separated lines both representing positive sentiment with different confidence scores in two sentences with both characters appearing in them.

Discussion

Conclusion

Throughout our rigorous exploration of character detection, sentiment estimation, and protagonist/antagonist detection, we ventured into a diverse range of methods, each yielding distinct degrees of success. Amidst this journey, we encountered common hurdles that frequently arise in the realm of natural language processing, including the intricacies of co-reference resolution and the accurate identification of named entities. It became evident that our implementation could be further refined by leveraging improved individual components, such as enhancing the performance of the named-entity recognition model or fine-tuning the scoring function for protagonist and antagonist detection.

It is important to acknowledge that our experimentation primarily revolved around fables, thereby leaving us with uncertainty regarding the performance of our solution when applied to lengthier texts. The transition to longer literary works would inevitably introduce new challenges that require careful consideration and tailored solutions. The complexities associated with maintaining coherence and extracting meaningful insights from extended narratives necessitate further investigation and adaptation of our existing techniques. Future endeavors will undoubtedly involve addressing the unique hurdles posed by more expansive literary compositions, thereby enhancing the robustness and applicability of our approach to diverse storytelling contexts.

References

[1] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, 2016.

[2] David Bamman, Sejal Popat, and Sheng Shen. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144, 2019.

[3] V Devisree and PC Reghu Raj. A hybrid approach to relationship extraction from stories. *Procedia Technology*, 24:1499–1506, 2016.

[4] 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.

[5] Anna Kazantseva and Stan Szpakowicz. Summarizing short stories. *Computational Linguistics*, 36(1):71–109, 2010.

[6] Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

[7] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[8] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[9] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

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

[11] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*, 2020.

[12] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. 2017.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[14] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87, 2023.