FRI Natural language processing course 2023
Project report

University of Ljubljana
Faculty of Computer and
Information Science

Literacy situation models knowledge base creation

Nina Velikajne, Jer Pelhan

Abstract

Keywords
sentiment analysis, named entity recognition

Advisors: Slavko Žitnik

Introduction

Literature is a rich and diverse field that contains a vast set of characters and relationships. From novels and short stories to plays and poems, literature offers an endless variety of human experiences, emotions, and perspectives. Each literary work presents a unique set of characters with their own personalities, backgrounds, and relationships with each other. These characters interact with each other in complex ways, creating intricate webs of relationships and social dynamics. Understanding these elements might be easy for a human, but it can present a difficult task for natural language processing (NLP) since natural language contains ambiguities and unclear references that might be difficult to understand for machines.

In this paper, we describe our approach for creating a knowledge base of literacy situation models using a combination of named entity recognition (NER), character graph co-occurrence, and character sentiment analysis. Our approach is based on the analysis of fairy tales that most often have very contrasting characters and more superficial sentence structures. By analysing character interactions in our knowledge base, we will gain insights into how their interactions shape the story, who is protagonist and antagonist of the story, and how the characters are portrayed.

The first step of our approach is applying NER to identify and classify different characters in fairy tales. The entities identified through NER techniques might be noisy, ambiguous, and lack clear semantics. Consequently, identifying connections between related entities in a given dataset, as well as with existing knowledge might enrich the data, and increase disambiguation and data consolidation. Dataset providers often seek to enhance a particular dataset by adding links to comprehensive reference datasets in order to achieve mentioned goals [1]. Herein, we plan to analyse the text to determine which characters interact with each other and how frequent their interactions are and construct a character graph co-occurrence graph. Then, we apply sentimental analysis algorithms to determine character morality and character-to-character sentiment. In the following, we will gain information about the morality of characters and from this in combination with character graph concurrence, conclude which characters are protagonists and antagonists, which hold side roles, etc. Furthermore, we plan to analyse the sentiment of whole stories, in order to label where bad or good events occur, and present them chronologically throughout the story.

Our corpus consists of 63 tales collected from [1]. Our method is used to automatically create a knowledge base on whole corpora. We evaluate the performance of our pipeline on a smaller subcorpora consisting of 10 tales, for which we write all the expected outputs, and basically annotate the data.

Related works

There are several approaches to NER in NLP. The simplest one are the rule-based systems. One of the first research papers in the field was presented by Lisa F. Rau [2]. It relies on heuristics and handcrafted rules. System was tested on over one million words and it successfully extracted company names with over 95% accuracy. Newer approaches use machine learning in order to classify NER. Mikheev *et al.* [3] proposed a supervised method based on maximum entropy. They report a recall of 76% for locations, 49% for organizations and 26% for persons with precision ranging from 70% to 90%. The techniques used in (NER) generally depend on lexical resources (such as WordNet), lexical patterns, and statistics derived from analyzing large unannotated corpora. Akbik *et al.* [4] presented contextual string embedding. They

[1] https://www.gutenberg.org/

use LSTM and Bi-LSTM language models and achieve F1 score of 93.09%. A similar approach is also used in the Stanza Python library, which we intend to use.

The analysis of character interactions is a critical aspect of NLP, and relation extraction plays a key role in this process. Lehmann *et al.* [1] developed RelFinder, a tool that utilizes a BFS algorithm to identify semantic connections among multiple entities in an RDF dataset. This algorithm is responsible for finding all related entities in the tripleset. Rather than focusing on finding connections between two given entities, paper [5] aims to identify the most connected entities in relation to a given relationship and entity. Although their approach offers an intriguing perspective on the problem, it differs from our problem. They search for connected entities based on known relationships, while we will try to uncover such connections between known entities. Leskovec *et al.* [6] presented a technique suggesting positive and negative relationships between people in a social network. This might also be useful in detecting the negative and positive characters in fairy tales.

Sentiment analysis is a widely researched topic used in commercial applications to predict user opinions on product reviews, detect harmful speech on social media, etc. Without using massive amounts of texts with the annotated sentiment, knowledge-based methods such as the method presented by Andreevskaia [7] rely on sentiment lexicons. Mentioned lexicons are lists of words that typically express sentiment, negative or positive, with the corresponding label (usually an integer), that represents positive to negative sentiment. But the performance of these methods heavily relies on each sentimental word being in the lexical base. Furthermore, some words in different contexts have different sentimental values, which cannot be distinguished using knowledge-based methods. Strapparava [8] a lexical database WordNet-Affect, that extends WordNet for sentiment analysis applications. WordNet-Affect provides a way to identify the emotional tone of the text by linking the words in the text to their corresponding affective information in the WordNet-Affect database. This way, the method can achieve more detailed information about the emotional meaning, ”happy” has a positive valence, even though the word ”mourning” is next to it in the method presented by Andreevskaia [7], but not with WordNet.

Methods

Corpus Analysis

We collected our corpus of Grimm fairy tales from Project Gutenberg [2], *i.e.*, free library of e-books. Our corpus is analysed in Table 1 and consists of 63 Grimm fairy tales, with an average length of approximately 1900 words. We find that on average the vocabulary size is 472 without any text preprocessing, which shows superficial sentence structures.

Table 1. Basic corpus analysis.

number of fairy tales	63
shortest (#words)	286
longest (#words)	4696
avg. num. of words	1915,17
avg. num. of sentences	44.57
avg. num. of words in sentence	42.97
vocabulary size	5723
avg. vocabulary size	472

Character Morality and Character-to-Character Sentiment analysis

For sentiment analysis of fairy tales, we use the VADER [9] rule-based approach for sentiment analysis. It uses a lexicon-based approach to determine the polarity (positive, neutral or negative) of the input text. For character-to-character sentiment analysis, we find all sentences and blocks of sentences where two characters interact with each other and analyse the polarity of the language towards each other. This way we determine their relationship and sentiment towards each other. For character morality analysis VADER is used to analyse the actions and words of a character in a story. By analysing the polarity of sentences that used by a character or sentences their name appears in, we portray the character as good or bad.

[2] https://www.gutenberg.org/

References

[1] Philipp Heim, Sebastian Hellmann, Jens Lehmann, Steffen Lohmann, and Timo Stegemann. Relfinder: Revealing relationships in rdf knowledge bases. *SAMT*, 5887:182–187, 2009.

[2] Lisa F Rau. Extracting company names from text. In *Proceedings the Seventh IEEE Conference on Artificial Intelligence Application*, pages 29–30. IEEE Computer Society, 1991.

[3] Andrei Mikheev, Marc Moens, and Claire Grover. Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8, 1999.

[4] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649, 2018.

[5] Yong-Jin Han, Seong-Bae Park, Sang-Jo Lee, Se Young Park, and Kweon Yang Kim. Ranking entities similar to an entity for a given relationship. In *PRICAI 2010: Trends in Artificial Intelligence: 11th Pacific Rim International Conference on Artificial Intelligence, Daegu, Korea, August 30–September 2, 2010. Proceedings 11*, pages 409–420. Springer, 2010.

[6] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social net-

works. In *Proceedings of the 19th international conference on World wide web*, pages 641–650, 2010.

[7] Alina Andreevskaia and Sabine Bergler. CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 117–120, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[8] Carlo Strapparava and Alessandro Valitutti. Wordnet-affect: an affective extension of wordnet. *Vol 4.*, 4, 01 2004.

[9] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.