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

Literacy situation models knowledge base creation

Nace Gorenc, Jernej Vrhunc in Žan Pečovnik

Abstract

Advisors: Slavko Žitnik, Aleš Žagar

Introduction

Natural language processing mainly focuses on different approaches of processing, analyzing and understanding large amounts of text. Extracting information from text can have great added value in many aspects of life such as science or economy. In recent years natural language processing became one of the hottest research topic. One of the driving forces was public availability of ChatGPT, which caused quite a stir since people started using it for various task which were formerly done by hand and with a little bit of help from search engines like Google. This report focuses on the analysis of personal relationships between fictional characters, using coreference resolution, named entity recognition (NER) and entity co-occurence graph (ECG).

Initial idea

In this assignment we will focus on creating a model, which will successfully recognize the relationships (positive, negative, neutral relation) between the characters in any short story. The initial idea is to collect sufficient amount of English short stories and annotate them in such a way that the model we are going to create, will be able to learn general relationships between characters from them. Because acquiring a large enough and exact amount of data is very important important in this kind of assignment, we will do this very precisely and on a large scale of books/stories.

Nextly our plan is to train a model on top of the dataset we created and see how it performs. With this results, we will obtain useful informations on how to improve the model, our dataset and how to possibly change the annotations we created for our dataset.

For extracting information about characters we will use named entity recognition (NER), which is a type of text processing which locates and classifies named entities into predefined categories such as *person*, *organization*, etc. We will use different NERs (Stanza [1], NLTK [2], Flair [3]) and compare the computed results to see which model prodcues the most accurate results.

Before extracting the information about entities from the text we will also perform coreference resolution which is a task of finding referring expressions (he, she, that, this, etc.) that refer to the same entity. For this task we will use the implementation from spaCy [4] and Allennlp [5] libraries to see which works best.

For detecting the actual relationship between characters we will construct a entity co-occurence graph (ECG) and find out from the graph, who interacts with whom and in which way.

Related work

A relation extraction boggled minds of the scientists long before NLP's renaissance driven by the invention of transformers.

In 2011, authors of [6] reviewed multiple supervised and semi-supervised classification approaches to the relation extraction task along with critical analysis. Among the supervised approaches, dependency path kernels stood out as the best both in term of computational complexity and performance. On the other hand, semi-supervised approaches seemed to be well suited for open domain relation extraction systems since they could easily scale with the database size and could extend to new relations easily.

Groza and Corde [7] focused on extracting information about literary characters from folktales. One of the tasks was to identify the main characters and the parts of the story where these characters are described or act. Their system relies on the folktale ontology which is based on Propp's model for folktales morphology and it defines three knowledge sources. The first one is *folktale morphology*, the second one is *folktale main entities* and the last one is *family relationships in folktale*. The authors used three sequential algorithms for retrieving information about characters. The main goals of

these algortihms were to extract characters from the folktales, use decoreference resolution and finding the perspective of the characters.

More recently the authors of [8] proposed method of a hybrid approach which combines the features of unsupervised and supervised learning methods, which also uses some rules to extract relationships. The method identifies the main characters and collects the sentences related to them. Then these sentences are analyzed and classified to extract relationships. For testing purposes, 100 short stories were used with around 300 character pair relations. This model identified parent-child relationships, friendships or if there was no relation.

In a recent survey [9], Vincent Labatut and Xavier Bost represented a character network, which can be used for extraction of relations between characters in works of fiction (e.g. novels, movies, plays, TV series). The main goal of this survey was to explain necessary steps, such as character identification, graph extraction and interaction detection, to construct such network.

Dataset analysis

Methods

Results

Future directions and ideas

Possible dataset: https://www.kaggle.com/datasets/leangab/poe-short-stories-corpuscsv?resource=download

Discussion

References

[1] Stanza - nlp package for many human languages. https://stanfordnlp.github.io/stanza/ner.html. Accessed: 19.3.2023.

[2] Natural language toolkit. https://www.nltk.org/. Accessed: 19.3.2023.

[3] Flair - a very simple framework for state-of-the-art nlp. https://github.com/flairNLP/flair. Accessed: 19.3.2023.

[4] Spacy - industrial-strength natural language processing. https://spacy.io/. Accessed: 19.3.2023.

[5] Allen nlp. https://allenai.org/allennlp. Accessed: 19.3.2023.

[6] Nguyen Bach and Sameer Badaskar. A review of relation extraction. 05 2011.

[7] Adrian Groza and Lidia Corde. Information retrieval in falktales using natural language processing. In *2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 59–66. IEEE, 2015.

[8] V. Devisree and P.C. Reghu Raj. A hybrid approach to relationship extraction from stories. *Procedia Technology*, 24:1499–1506, 2016. International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015).

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