FRI Natural language processing course 2023
Project report

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

Character analysis in Slovene short stories

Gal Petkovšek, Marija Marolt, Jana Štremfelj

Abstract

The abstract goes here.

Keywords
Natural Language Processing, character analysis, knowledge base, character relations

Advisors: Slavko Žitnik

Introduction

Natural Language Processing (NLP) is an evolving field trying to automatically extract the knowledge from the abundant volume of natural language text. Machines can process significantly larger amounts of text, but there are some challenges such as ambiguity of the text, the sentiment analysis, commonsense reasoning, information extraction etc.

The main task of our project is the knowledge base creation, which is a sub-field of Natural Language Processing. Based on this knowledge base the relations between the characters are created. Each character is also classified as positive, negative or neutral and the protagonist of the story is defined.

The dataset that we have chosen, is a collection of Slovene short stories. Our training corpus containes 65 stories writen by Ivan Cankar. They are abtained from dLib.si, which is a digital library and document server for Slovenian language resources.

The pipeline and models are adjusted to Slovene language. We get a graph from the information extraction pipeline in which the nodes represent the characters and links the relations between them. First step in the pipeline is coreference resolution where all expressions that refer to the same entity are found, in next step the recognition of all the entities is made (e.g. person, location, organization... however we only focus on characters). In the last step the knowledge graph is constructed [1].

The visualization is going to represent the relations of the characters and their profiles, which we obtain from the knowledge graph.

Related work

Character analysis and relationship extraction is already a well researched topic that falls under the category of knowledge base extraction from text. Different authors report different approaches to tackling the problem of knowledge base creation, while most of them still roughly separate the pipeline steps to the ones meant to extract nodes ans the ones to extract edges. In an article by Kertkeidkachorn *et. al.* [2] they describe a pipeline for knowledge graph construction using with five steps: entity recognition (and mapping), coreference resolution, triple extraction, triple integration and predicate mapping. In another article [3] the authors describe the an algorithm called Grapher that uses an encoder-decoder language model (sequence to sequence algorithm) for node extraction and for edge detection they tested two different approaches (LSTM/GRU and classification).

When we consider the character analysis problem (which is a sub-problem of the above described knowledge base creation problem). Different research is also being done in that area. The article [4] by Nalisnick *et. al.* for example predicts the relationships from characters on Shakespeare's plays, however their problem is different from ours in the way that since we have plays the problem is mainly focused on the dialog between characters. Another example of character analysis is described in the article [5] where the author's main focus is to classify the character as positive or negative which is done in 3 different ways: character's speech, character's actions and predicatives.

In slovene however, the topic is not as well covered. Markovič [6] *et. al.* in his article for example uses network analysis for analysing Slovene fables, however it does not strictly cover sentiment analysis of characters. In another article [7] the authors describe how they prepared a Slovene database for the problem of classifying texts as positive, negative or neutral, however they do not focus on the characters in those texts. Therefore to the best of our knowledge not many authors have attempted to solve the character analysis problem purely in slovene language.

In our work we will be using the tools and models for slovene language edveloped undet the project Razvoj slovenščine v digitalnem okolju [8] which offers very modern models for solving differen NLP tasks.

Methods

Results

Discussion

Acknowledgments

References

[1]

[2] Natthawut Kertkeidkachorn and Ryutaro Ichise. T2kg: An end-to-end system for creating knowledge graph from unstructured text. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[3] Igor Melnyk, Pierre Dognin, and Payel Das. Knowledge graph generation from text. *arXiv preprint arXiv:2211.10511*, 2022.

[4] Eric T Nalisnick and Henry S Baird. Character-to-character sentiment analysis in shakespeare's plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 479–483, 2013.

[5] Lucie Flekova and Iryna Gurevych. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816, 2015.

[6] Rene Markovič, Marko Gosak, Matjaž Perc, Marko Marhl, and Vladimir Grubelnik. Applying network theory to fables: complexity in Slovene belles-lettres for different age groups. *Journal of Complex Networks*, 7(1):114–127, 08 2018.

[7] Jože Bučar, Martin Žnidaršič, and Janez Povh. Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, 52:895–919, 2018.

[8]