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

Literacy situation models knowledge base creation

Anže Mur, Jaka Bernard, Edo Ljubijankić

Abstract

Keywords
content-based analysis

Advisors: Slavko Žitnik

Introduction

Literature is almost infinitely diverse and as such contains many characters, events, and variations on those. During education, young people are often made to understand the relationships and interplay amongst various characters, factions, and events in numerous literary works. This is usually not very difficult for people but can be a hard nut to crack for computers, as natural language contains ambiguities and unclear references that are difficult to understand for machines, but possible to parse apart from context clues for human readers.

Our code is publicly available on GitHub [1].

Related work

The authors of [2] present a new corpus of 50k crowdfunded five-sentence commonsense stories called ROCStories. This corpus is used in their evaluation framework Story Cloze Test which requires a system to select an appropriate right or wrong ending for the four-sentence tested story. The presented framework can serve as a generic story understanding and story generation evaluation framework. Another evaluation benchmark corpus is presented in the article [3] with the focus on the evaluation of reconstruction of event-driven plot structures.

Authors of [4] present an ACE entity annotated dataset, collected from 100 different English literary texts obtained from Project Gutenberg. Corpus contains 210,532 tokens that are categorized into six different categories - person, organization, location, geopolitical entity, facility, organization, and vehicle. Additional work on the corpus was done in [5] which authors collected the annotations for the events. They also presented and evaluated the current SOTA models on the collected corpus.

Knowing relations between characters is an important part of story understanding and summarization. The authors of the article [6] present a method that identifies the main characters in a story and extracts relations between them. The proposed method is a hybrid approach that combines the features of unsupervised and supervised learning methods. For testing purposes, 100 short stories for kids were used and 300 character pair relations analyzed. The system identified if the relation between pair of characters is 'Parent-Child' or 'Friendship' or if there is no relation found. The proposed method proved to be good compared to other existing methods.

In article [7] authors present linguistically informed deep neural network architecture for automatic extraction of cause-effect relations from text documents. The proposed architecture uses word-level embeddings and other linguistic features for detecting such events. Bi-directional LSTM and other models were used and compared in experiments. For testing, multiple datasets were used. Most tests were run on SemEval 2010 annotated dataset, ADE or adverse drug effect dataset, BBC News Article dataset and Recall dataset. By understanding how the neural network works, it could be applied to extract cause-effect relations from short stories.

Dataset

For the dataset, we used books available at Project Gutenberg [8], which provides over 60 thousand free eBooks digitized by volunteers. The project is focused on older work, for which the U.S. copyright has expired. To obtain the dataset we scrapped the website for English short stories with a public license. We focused on the books that are available in text format and so we obtained 818 short stories.

This collected dataset is not annotated so we could annotate it ourselves, by collecting all of the annotations for each of the ACE entities from the corpus presented in [4] since the corpus is collected from the same origin. We could then

map collected annotations to our dataset by mapping tokens one to one - we can assume that dataset would be only partly annotated. Another option is to take an intersection of short stories between our collected stories and the annotated corpus.

Methods

Our initial ideas include:

- **Identification and analysis of the events** - We want to identify the events occurring in the story. Additionally, we want to collect more information around these events: participants, occurring actions, time and space components. Based on this information we can estimate the importance of the identified events and form causal relationships between them.
- **Identification and analysis of the characters** - We want to identify the characters/participants and their roles and importance in the story (ex. antagonist or protagonist).

References

[1] Literacy situation models knowledge base creation. https://github.com/anzemur/literacy-knowledge-base. Accessed: 18-03-2022.

[2] 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.

[3] Tommaso Caselli and Piek Vossen. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, 2017.

[4] 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.

[5] Matthew Sims, Jong Ho Park, and David Bamman. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, 2019.

[6] 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).

[7] 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, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] Project gutenberg. https://www.gutenberg.org/. Accessed: 18-03-2022.