FRI Natural language processing course 2021
Project report

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 one of the datasets, we use the seven stories provided by the teaching assistant for this topic. While not annotated, we were able to obtain the text and convert it into word graphs, which we were able to process further. We also obtained the climaxes for the seven stories by hand from Google.

Another dataset was the EventStoryLine corpus [8], a collection of short internet articles, annotated by both experts and community members alike to include some temporal markings, the climax, and in some cases causality between events. Sadly the graphs obtained from parsing these annotated articles were very sparsely connected, giving us little to work with.

For the third dataset, we used books available at Project Gutenberg [9], 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.

Methods

0.1 Attempts at causality and climax detection

Our first attempt at causality detection involved the EventStoryLine corpus [8]. We attempted to construct graphs of words and perform analysis on the graphs to extract causality and climaxes. Sadly, the entries in the corpus were short articles and the format was very sparsely connected, so not much could be done in terms of network analysis.

Next we looked into causality detection attempts of others [10, 11, 12, 13, 14, 15, 16]. We have found that most attempts focus on detecting causality within a single sentence or short piece of text, such as an article introduction, making it less than ideal for detecting causality in longer stories.

0.2 Story graph analysis

For initial parsing and processing we used the seven stories provided by the teaching assistant. We used the spacy and torch geometric packages to convert the stories into graphs. From those graphs, we extracted the largest connected subgraph, as well as relabeling the nodes in the original network from nodes with word labels to nodes that are words, losing structure but gaining connections. We then extracted the largest connected sub-graph from this graph as well, netting us a total of four graphs per story. An example visualization of an extracted largest connected sub-graph can be seen in Figure 1, where the visualization represents the text ”*Absolutely, the only terms on which you can have him restored to you are these: We demand fifteen hundred dollars in large bills for his return; the money to be left at midnight to-night at the same spot and in the same box as your reply as hereinafter described.*” within the story *The Ransom of Red Chief*, by O. Henry

We computed a variety of scores for each node in the above graphs, including node degree, harmonic mean distance, and PageRank, then obtained the top 10 nodes for each score in each graph and printed them to a file. After obtaining the climax descriptions for the seven stories by hand, we compared the described climaxes to our results. Sadly, only one story's climax appears significantly in the results, and likely because of many repetitions of the word ”please” in the climax of the story, which affects the scores in the graph.

0.3 Sentiment analysis of character relationships

We wanted to extract list of important characters in a story and tried to find if relationship between pair of characters is a positive or negative. In work [6] authors assume that two characters who appear in a sentence together have some sort of relationship. In our approach we extract characters and perform sentiment analysis, calculate sentiment matrix and visualize relationships with graph.

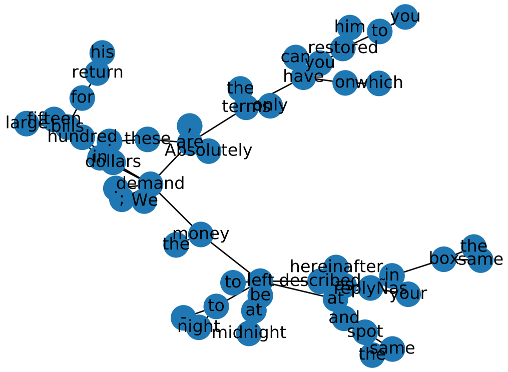

Figure 1. A story word graph visualization. This is a visualization of the largest connected sub-graph of the story The Ransom of Red Chief, by O. Henry.

We first got the list of characters from text with Name Entity Recognition model. For this we used pre-trained Spacy NER classifier. We also count number of occurrence of each character in story. The most frequently occurred characters seem to have larger role. We then used NLTK library to split text into sentences. We use library Afinn to calculate the sentiment score of each sentence. Sentence with positive words like ”love”, ”good”, ”smile” etc. leads to higher sentiment score and implies that the characters that co-occur in such context are likely to be in positive relationship. With the gathered data we calculate sentiment matrix. A larger sentiment score value makes relationship between pair of characters more positive (friends, lovers etc.) while lower or negative value of sentiment score indicates there is a negative relationship (enemies, rivals etc.).

We output sentiment graph where each node represents a character in the story and each edge represents the relationship between the two characters it's connected to. Node size represents the importance of a character. Each edge has a different color for better understanding of relationships. The brighter color of the edge indicates more friendly relationship while darker color of edge shows more hostile relationship. Example of graph for a story is shown in figure 2.

Future directions and ideas

Another avenue for climax and event detection from stories would be to use the above graphs and try to teach some neural networks to detect our desired events. For this we could use the steps provided during labs, although labeling could pose quite a challenge, especially if we expend to more than seven stories.

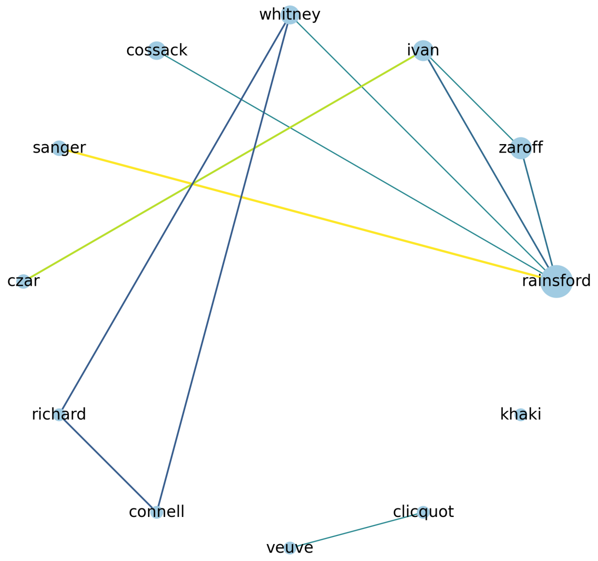

Figure 2. Sentiment graph for story The Most Dangerous Game.

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] Eventstoryline. https://github.com/tommasoc80/EventStoryLine. Accessed: 03-04-2022.

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

[10] Ning An, Yongbo Xiao, Jing Yuan, Jiaoyun Yang, and Gil Alterovitz. Extracting causal relations from the literature with word vector mapping. *Computers in biology and medicine*, 115:103524, 2019.

[11] Xianxian Jin, Xinzhi Wang, Xiangfeng Luo, Subin Huang, and Shengwei Gu. Inter-sentence and implicit causality extraction from chinese corpus. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 739–751. Springer, 2020.

[12] Nabiha Asghar. Automatic extraction of causal relations from natural language texts: a comprehensive survey. *arXiv preprint arXiv:1605.07895*, 2016.

[13] V Khetan, RR Ramnani, M Anand, S Sengupta, and AE Fano. Causal-bert: Language models for causality detection between events expressed in text, corr abs/2012.05453 (2020). *URL: https://arxiv.org/abs/2012.05453*, 2012.

[14] Abbas Akkasi and Mari-Francine Moens. Causal relationship extraction from biomedical text using deep neural models: A comprehensive survey. *Journal of Biomedical Informatics*, 119:103820, 2021.

[15] Yifan Shao, Haoru Li, Jinghang Gu, Longhua Qian, and Guodong Zhou. Extraction of causal relations based on sbel and bert model. *Database*, 2021, 2021.

[16] 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.