FRI Natural language processing course 2022/2023
Inital report

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

Literacy situation models knowledge base creation

Jaša Frelih Črv, Jurij Kolenik, and Žan Jonke

Abstract

Keywords
Story entities, Family relationship extraction, Short stories

Advisors: Slavko Žitnik

Introduction

Every story we read contains a lot of information that can be extracted using literacy situation models. We can extract different characters, how they are connected between each other and the general sentiment about them. We can go even further by looking at events in the story and analyzing the space and time around them. By analyzing the sentiments, co-occurrence, and importance of characters in a story, we can quickly grasp the content and meaning of the story as a whole. This is especially useful for people who may not have a lot of time to read or analyze lengthy texts.

For our project, we will focus on character analysis and sentiment analysis of these characters. We will build a pipeline, which will perform the tasks listed below:

- Named entity recognition (NER) to extract characters.
- Character sentiment analysis to determine whether a character is good or bad.
- Build a character co-occurrence graph to see which characters interact.
- Calculate entity importance score.

For NER, we will use the Stanza package, which contains many tools for linguistic analysis. To determine whether a character is good or bad, we will look at the general sentence occurrence sentiment, where we will analyze the general theme of sentences that a character occurs in and determine the sentiment based on the average context of the sentences. For character co-occurrence graph and entity importance score, we will look at different methods. Our corpus will contain the stories which were provided to us in the project description, and we will add additional short and medium length stories.

1. Existing Solutions & Related work

When we started working on this project, we firstly looked at some existing papers that have the same topic or are connected to the same field of research as ours.

The problem of automatic sentiment analysis (SA) is a growing research topic. Although SA is an important area and already has a wide range of applications, it clearly is not a straightforward task and has many challenges related to natural language processing (NLP) [1]. The simple idea behind computing an emotional figure profile is that the strength of semantic associations between a character (name) and the prototypical "emotion words" contained in the label list gives us an estimate of their emotion profile [2]. In article [3] character sentiment analysis is conducted using a simple word-list based approach where the "emotion" of characters such as "Miss Elizabeth Bennet" from Jane Austen's Pride and Prejudice by counting all emotional words in paragraphs that featured only one character and adding them to the character's total. Further building on top of this, [4] exploited text structure and knowledge-based SA to track the emotional trajectories of interpersonal relationships rather than of a whole text or an isolated character. To extract these relationships, they mined for character-to-character sentiment by summing the valence values (provided by the AFINN sentiment lexicon [5]) over each instance of continuous speech and then assumed that sentiment was directed towards the character that spoke immediately before the current speaker. This assumption however does not always hold, as a conversation between two characters might be centered around expressing sentiment about someone offstage.

In the article [6], the authors examined the network of named entity co-occurrences in written texts, which coincides directly with our proposed work on our project. They pointed out that the use of methods borrowed from statistics and physics to analyze written texts has allowed the discovery of unprecedented patterns of human behavior and cognition by establishing links between models features and language structure. While current models have been useful to unveil patterns via analysis of syntactical and semantic networks, only a few works have probed the relevance of investigating the structure arising from the relationship between relevant entities such as characters, locations and organizations. They focused on the representation of entities appearing in the same context as a co-occurrence network, where links are established according to a null model based on random, shuffled texts in their paperwork. With some simulations performed in novels, they discovered that their proposed model displayed interesting topological features, such as the small world feature, characterized by high values of clustering coefficient. The effectiveness of their model was later also verified in a practical pattern recognition task in real networks. When they compared their own model with traditional word adjacency networks, it displayed optimized results in identifying unknown references in texts. They concluded that the proposed representation plays a complementary role in characterizing unstructured documents via topological analysis of named entities. Their finally thought that their method could be used to improve the characterization of written texts (and related systems), especially if it was combined with traditional approaches based on statistical and deeper paradigms.

The next article [7] focused on named Entity Recognition. Named Entity Recognition serves as the basis for many other areas in Information Management. However, it is unclear what the meaning of Named Entity is, and yet there is a general belief that Named Entity Recognition is a solved task. The authors of this paper we analyzed the evolution of the field from a theoretical and practical point of view. They argued that the task is actually far from solved, and also presented the consequences for the development and evaluation of tools. They also discussed topics for further research with the goal of bringing the task back to the research scenario. We also believe that some ideas that the authors established in their work could be beneficial for our research work.

The article [8] focuses on data-driven relation discovery from unstructured texts. The authors proposed a data driven methodology for the extraction of subject-verb-object triplets from a text corpus. Previous works on the field solved the problem by means of complex learning algorithms requiring hand-crafted examples. They stated that their proposal completely avoids learning triplets from a data-set, and they built it on top of a well-known baseline algorithm. The baseline algorithm uses only syntactic information for generating triplets and is characterized by a very low precision, meaning that very few triplets are meaningful. Their idea is to integrate the semantics of the words with the aim of filtering out the wrong triplets, thus increasing the overall precision of the system. The algorithm has been thoroughly tested and has shown good performance with respect to the baseline algorithm for triplet extraction. We also believe some ideas presented in this paper to be useful during our own research work.

In the paper that our assistant recommended [9] the authors focused on automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks. The author proposed a linguistically informed recursive neural network architecture for automatic extraction of cause-effect relations from text. These relations can be expressed in arbitrarily complex ways. The architecture they used relies on word level embedding and other linguistic features to detect causal events and their effects mentioned within a sentence. They used the extracted events and their relations to build a causal-graph after clustering and appropriate generalization, which is then used for predictive purposes. They have evaluated the performance of the proposed extraction model with respect to two baseline systems. Furthermore, they also compared their final results with related work reported in the past by other authors on *SEMEVAL* data set, and found that their proposed model works better.

References

[1] DMilind Soam and Sanjeev Thakur. Sentiment analysis using deep learning: A comparative study. In *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)*, pages 1–6, 2022.

[2] Arthur M. Jacobs. Sentiment analysis for words and fiction characters from the perspective of computational (neuro-) poetics. Frontiers in Robotics and AI, 2019.

[3] Micha Elsner. Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644, Avignon, France, April 2012. Association for Computational Linguistics.

[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, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[5] Finn Årup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. 2011.

[6] Diego Raphael Amancio. Network analysis of named entity co-occurrences in written texts. *Europhysics Letters*, 114(5):58005, jul 2016.

[7] Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards Interfaces*, 35(5):482–489, 2013.

[8] Marilena Ditta, Fabrizio Milazzo, Valentina Raví, Giovanni Pilato, and Agnese Augello. Data-driven relation discovery from unstructured texts. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, volume 01, pages 597–602, 2015.

[9] 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.