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

Literacy situation models knowledge base creation

Mila Marinković, Ilina Kirovska and Radoslav Atanasoski

Abstract

Introduction

Short stories are powerful mediums that in spite of their length often convey complex themes and ideas. Understanding them demands a thorough grasp of the context in which they were written. The purpose of our project is to help the readers in bettering their understanding and knowledge of the stories by building a knowledge base. We can build the knowledge base by extracting different types of data and by doing that we are able to provide different kinds of story information to the user. The initial idea is to concentrate on the analysis of the relationships between characters and possibly expanding our results by determining the genre of a particular story with the help of the events(temporal information extraction) and the setting of the story.

Related work

In paper[1], the supervised and unsupervised approaches are integrated to produce a hybrid model that can extract relationships from stories.The model identifies the main characters, gathers sentences that are relevant to them, analyzes these words, and then assigns a classification to the characters' relationships. If a sentence is classified during training, it passes through the supervised section of the model; otherwise, it goes through the unsupervised section. The corpus utilized was a collection of 100 children's short stories that featured a variety of different relationship types. This model identified if a relationship between characters was "Parent-Child", "Friendship", or "No-Relationship". When compared to other approaches, the hybrid model performed well.

A rule-based approach to character identification and family connection analysis is presented in paper[2]. The authors used fundamental NLP procedures such as tokenization, POS tagging and sentence parsing to extract the entities and relationships among them. There are two key modules i.e. Character extraction and Relationship extraction. For the character extraction, the NER kit provided by NLTK, was used to determine the possible characters. Ultimately, from all results, the entities labeled as 'persons' were identified as characters. Regarding the relationship extraction parse tree generation was used where each entity was given a few properties including relation id. Following the pronoun resolution and character extraction, the relation id of each entity is modified during the execution of relation resolution on each parse tree. After the identification of the some relationships, new relationships are resolved using the propagation rules for relationship. Considering that character extraction accuracy equaled to 95.33% and relationship extraction accuracy to 80.5% it is evident that this approach was effective.

References

[1] 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).

[2] Alisha Bajracharya, Saurav Shrestha, Sharmila Upadhyaya, Bk Shrawan, and Subama Shakya. Automated characters recognition and family relationship extraction from stories. In *2018 8th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pages 14–15, 2018.