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

Challenges in Creating a Knowledge Base for Literacy Situations: Character Extraction and Analysis

Matic Šuc and Loris Štrosar Grmek

Abstract

Keywords
named entity recognition, sentiment analysis

Advisors: Slavko Žitnik

Introduction

Natural Language Processing (NLP) is a subfield of Artificial Intelligence (AI) that focuses on enabling computers to understand and generate human language. One of the main challenges in NLP is to build systems that can reason about text in a way that mimics human cognitive abilities. To achieve this goal, researchers have developed a variety of models that capture different aspects of language understanding, such as syntax, semantics, and pragmatics. One such model is the Literacy Situation Model (LSM), which captures the knowledge and assumptions that people use to understand written text. The LSM is based on the idea that readers construct mental models of the situations described in texts, and that these models help them to interpret and remember the information conveyed. Creating a knowledge base for LSMs involves extracting and organizing the relevant information from texts, such as entities, events, and relationships, and encoding it in a structured format that can be used for reasoning and inference. This process involves a range of NLP techniques, such as named entity recognition, coreference resolution, and semantic role labeling. The resulting knowledge base can be used for a variety of applications, such as text understanding, question answering, and summarization.

Related work

In [1] a novel method for automatically extracting relationships between characters in a story using a hybrid approach that combines rule-based and machine learning techniques is proposed. The authors present a detailed analysis of their proposed method and evaluate its performance using a dataset of short stories. The results of their experiments show that their approach outperforms existing methods for relationship extraction from stories.

Another notable article [2] proposes a method for automatically extracting causal relationships from text using deep neural networks that are informed by linguistic features. The authors evaluate their method on a dataset of news articles and demonstrate its effectiveness in identifying causal relationships between events. This article showcases the potential of deep learning approaches for extracting meaningful information from unstructured text data and was presented at a conference on discourse and dialogue.

Methods

Coreference Resolution - CR

We used **AllenNLP** library, which is a NLP toolkit for building and evaluating deep learning models. Coreference resolution is the task of identifying all the expressions in a text that refer to the same real-world entity, such as ”he” referring to a previously mentioned person. The specific model being used is called ”coref-spanbert-large,” which is a large neural network model trained on a massive amount of text data. The model is designed to identify coreferences across long documents, making it well-suited for large-scale natural language processing tasks. By using this model, the NLP application can better understand the relationships between entities and improve the accuracy of downstream tasks such as question answering and summarization.

Named-Entity Recognition (NER)

We tried several NER libraries including:

- The NLTK (Natural Language Toolkit) is a popular

Python library used for NLP tasks such as NER, and it provides various tools and resources for processing natural language text.

- Spacy - a popular open-source NLP library that provides efficient and accurate Named Entity Recognition capabilities for various languages.
- Stanza - a state-of-the-art open-source NLP library that offers robust and multilingual Named Entity Recognition capabilities using deep learning models.

Character sentiment analysis

- SiEBERT - a fine-tuned checkpoint of RoBERTa-large [3]
- Stanza's pipeline

Graph visualisation

Visualize the sentiment between characters.

Results

Model	Precision	Recall	F Measure
Stanza	0.62	0.6	0.61
NLTK	0.6	0.21	0.31
Spacy	0.8	0.57	0.66

Table 1. Results NER models.

Model	Detected protagonists	Detected antagonists	All
SiEBERT	33	23	55

Table 2. Initial results of sentiment analysis.

Discussion

Acknowledgments

References

[1] V Devisree and PC Reghu Raj. A hybrid approach to relationship extraction from stories. *Procedia Technology*, 24:1499–1506, 2016.

[2] 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.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.