**UL FRI Data Science - Natural Language Processing, 2021-2022**
**Interim report**

# Literacy situation models knowledge base creation

Uroš Škrjanc[1], Matthew Tanti[2], and Marko Novak[3]

**Abstract**

**Keywords**
story entities, family relationship extraction, short stories

[1] *us1883@student.uni-lj.si, 63030323*
[2] *mt9734@student.uni-lj.si, 63210511*
[3] *mn3983@student.uni-lj.si, 63130166*

## Introduction

The field of recognising the content and structure of texts and extracting content from them is increasingly related to machine learning methods. Not only syntax checking but also pure understanding of texts by humans is interesting for machine learning for different motives. There is practically no more field related to language that in some way could not be linked to machine learning methods.

## Methods

In the group, we decided to try to analyse different short stories to extract information about the characters who appear in the texts and to find out what kind of relationships these characters are in.

First, we found a collection of English short stories and chose a subset of them for training. Then, in the python programming language, we will use libraries such as NLTK, SpaCy, SNER, GATE... to determine which characters appear in the stories. At this stage, it will be necessary to extract the characters' actual names from the text and ensure that the correct numbers of characters and their positions in the text are extracted.

In the second part, we will try to make a model that properly connects characters in pairs into family relations. The model should determine whether a couple of characters are in a family relationship and, if so, then determine the type of relation. Due to the complexity of the task, we decided not to classify relations which are not family-related. We think that discovering family relations is a complex enough task.

As part of the task, we will review the models that have already been implemented and, based on the results of the solutions already made, make our own model that would classify as accurately as possible.

## Existing solutions

The use of deep neural networks is on the rise in already implemented solutions, but other approaches and combinations of these are also used in models for classifying family relations between persons in the text, such as rule-based approaches, utterance attribution and vocative detection technique and unsupervised approach to the extraction of interpersonal relations are described in the articles, which we intend to take as a basis for further work.

## Relations

We decided to check if the following relations occur between the character in short stories:

- Wife
- Husband
- Mother
- Father
- Daughter
- Son
- Sister
- Brother
- Grandmother
- Grandfather
- Granddaughter
- Grandson
- Mother-in-law
- Father-in-law
- Daughter-in-law

- Son-in-law
- Sister-in-law
- Brother-in-law
- Aunt
- Uncle
- Niece
- Nephew
- Cousin

Those are the broadest family relations we intend to use. We expect that there will be no need to add new relationships. If the model proves to be too large for our task, some relations can be removed from the model.

## Model

For extracting and classification family relatins in short stories, we are considering two models: CasRel [1] and DocuNetl [2].

CasRel is sentence oriented. That means that in the first step model discovers identities and in the second step applies relation-specific taggers to identify all possible relations and the corresponding object simultaneously. Relations are represented with relational triples (subject, relation, object). The model takes all subjects in a sentence and, for each subject, iterates through relations and tries to find the object which is in a known relation with the subject. The model implements a more holistic approach to relation classification than usual, and, according to article [1], it achieved high accuracy on NYT and WebNLG datasets.

On the other hand, DocuNET (Document U-shaped Network) is document-oriented. That means the model tries to find relations between entities over the whole document. According to the article [2], above 40,7% of relations can only be identified at the document level, and this model is the first that implements relation extraction as semantic segmentation. The approach is similar to computer vision, where each pixel is classified into one visual object. On the same principle, DocuNET tries to classify two entities into one of the known relations.

We think that the second model would be more appropriate for our task because of the document level orientation. However, we're trying to use the first model to aid in annotating the data in preparation for training the final model since annotating all the data by hand would take a lot of time and effort.

## References

[1] Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, 2020.

[2] Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. Document-level relation extraction as semantic segmentation. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.