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

Literacy situation models knowledge base creation

Jan Krivec, Matija Gerčer and Rok Bosil

Abstract

Our selected project is Literacy situation models knowledge base creation. Here we will introduce the initial ideas about the specific topics of the selected project and existing related work done on these topics.

Keywords
Natural language processing, Knowledge base

Advisors: Slavko Žitnik

Introduction

The field of natural language processing largely focusses on approaches of processing, analyzing and understanding large amounts of text. Namely, we can extract important information from texts that can have a great added value for both science and the economy. In this project, we will focus on models for building a knowledge base. More precisely, we will create a model that determines the entities (determine the main characters) and relations between them from short stories, and we will present the results on a knowledge graph and try to obtain some useful information from it.

We will first focus on data cleaning and preprocessing, and then we will present some (semantic and syntactic) techniques for entity and relation extraction. Much of the attention will be paid to entity resolution, where we will identify whether multiple mentions in a text refer to the same entity. Finally, we will focus on the evaluation of the developed model by comparing it with existing examples.

Related work

In [1] the authors focused on extraction of information about literacy characters. One of their tasks was to indentify and classify the main characters in folktales. They extracted knowledge about characters by combining natural language processing and reasoning on ontologies. They described three algorithms for identifiying knowledge about charcters.

In [3] the authors propose a hybrid method of main character indetification and extraction of the relationships between them. This approach mainly combines supervised and unsupervised learning methods. In contrary to unsuperised learning, supervised learning algorithms learn to classify with the help of already known realtion types between entities

Methodology

In the first part, we will focus on cleaining and pre-processing the thext data with techniques such as tokenization, stemming, lemmatization, lower casing, removing stop words punctuations etc.. Especially we will focus on entity resolution, where we will identify whether multiple mentions in a text refer to the same entity. Then we will divide the development of the model into two parts, namely:

1. **Entity extraction:** Our algorithm will first determine all the entities in the selected short stories. We aim to further differenciate between different types of entities, for example inscances of people, locations, objects and so on. We will primarely focus on people and their relations. Here we can use sequence labeling, deep lerning models or rule-based approaches.

2. **Relation extraction:** This extends the entity extraction portion of the task. Here we will determine retations between entities, more specifically between characters. This task can be done with the use of syntactic patterns, suprevised machine lerning or unsuprevised machine learning.

   For enitiy extraction we wil use named-entity recognition.

   For relation extraction there are multiple models that can be used, such as MLPs, gradient boosting, Word2Vec and more. We can also use end-to-end models, such as BERT-based, CNN-based or RNN-based precomputed models [2].

Corpus analysis

We gathered 42 short stories. Stories have on average 2541,7 tokens, if we excluded punctuation marks. Each word has an average of about 4 characters. The average cosine similarity between short stories is 0,58.

We also included some slovene stories but corpus analysis was done only on english short stories.

References

[1] Adrian Groza and Lidia Corde. Information retrieval in folktales using natural language processing, 2015.

[2] Sebastian Ruder. http://nlpprogress.com/english/relationship_extraction.html.

[3] 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).

[4] Natalia Konstantinova. Review of relation extraction methods: What is new out there? In *Analysis of Images, Social Networks and Texts*, pages 15–28, Cham, 2014. Springer International Publishing.

[5] Yifan Peng, Manabu Torii, Cathy H. Wu, and K. Vijay-Shanker. A generalizable nlp framework for fast development of pattern-based biomedical relation extraction systems. *BMC Bioinformatics*, 15(1):285, Aug 2014.

[6] Thien Huu Nguyen and Ralph Grishman. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st workshop on vector space modeling for natural language processing*, pages 39–48, 2015.

[7] Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing*, 257:59–66, 2017. Machine Learning and Signal Processing for Big Multimedia Analysis.

[8] Jinhua Dou, Jingyan Qin, Zanxia Jin, and Zhuang Li. Knowledge graph based on domain ontology and natural language processing technology for chinese intangible cultural heritage. *Journal of Visual Languages Computing*, 48:19–28, 2018.

[9] Guanying Wang, Wen Zhang, Ruoxu Wang, Yalin Zhou, Xi Chen, Wei Zhang, Hai Zhu, and Huajun Chen. Label-free distant supervision for relation extraction via knowledge graph embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2246–2255, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[10] Rolf A. Zwaan. Situation models: The mental leap into imagined worlds. *Current Directions in Psychological Science*, 8(1):15–18, 1999.

[11] Rolf A. Zwaan, Mark C. Langston, and Arthur C. Graesser. The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, 6(5):292–297, 1995.

[12] Danielle McNamara and Arthur C. Graesser. *Coh-Metrix: An automated tool for theoretical and applied natural language processing*, pages 188–205. IGI Global, 2011.

[13] 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.

[14] 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, San Diego, California, June 2016. Association for Computational Linguistics.

[15] Tom Trabasso and Paul van den Broek. Causal thinking and the representation of narrative events. *Journal of Memory and Language*, 24(5):612–630, 1985.

[16] Shingo Nahatame. Revisiting second language readers' memory for narrative texts: The role of causal and semantic text relations. *Reading Psychology*, 41(8):753–777, 2020.