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 some well-known novels, and we will present the results on a knowledge graph and try to obtain some useful information from it, such as whether two people are in a friendly or hostile relationship and finding out how relationships change through chapters of the book. 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 extraction of entity relations using a variety of techniques and models. Finally, we will focus on the evaluation of the developed model by comparing it with existing prior knowledge from various sources (Wikipedia, Fandom, web web scraping).

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

Experimentation

Our goal was finding relations between characters in Game of Thrones books and try to determine their relation. We followed the following pipeline:

1. **Preprocessing** First we preprocessed the text by removing all stop words and punctuation marks.We also used a list of common English words to help our model find the correct entities.

2. **Entity extraction:** We used named entity recognition (NER) to identify all the character names showed up in the nove, which we later used for processing co-occurrence between them and trying to find the sentiment score. We used library Spacy for finding all people using named entity recognition, where we limited ourself to only entities of type *PERSON*. Because of Spacy library uses a lot of memory, we ran NER on each sentence as opposed to running it through the whole book. We also shortened all the names to only first name, to try to avoid repetitions. To try to minimise wrong character identification, we used a list of common words and filtered out the recognised entities, that showed up in the list of common words. After finding all the entities, we went through all the characters and counted all the occurencies (for this we used CountVectorizer from Sickit-Learn text). For further analysis we only used characters, that were the most mentioned.

3. **Relation extraction using sentiment analysis:** After finding all the characters we focused on finding the relations between them. First we had to compute the co-occurence matrix to find in which sentences multiple characters are mentioned. To calculate co-occurrence, we first need a binary occurrence matrix, that gives information on whether a name occurs in each sentence. Then, the co-occurrence matrix equals the dot product of occurrence matrix and its transpose. And we only take the lower triangle of the co-occurence matrix because it is simetrical through the diagonal.

   After finding the co-occurence matrix, we needed to add sentiment, to determine the relations between the characters. We used NLP library Afinn to calculate sentiment for each sentence. We used the results from the sentiment analysis and the co-occurence matrix, to find the final weights for each character pair.

After finding all the entities and determinig relations between them, we used the networkx library to plot the graph. We used the sentiment matrix for the weights on the edges of the graph. We tried runnign our algorithm on the first book of Song of Ice and Fire: A Game of Thrones. The results can be seen in figure 1. Lighter colors on edges represent friends or allies and darker colors represent enemies.

As we can see from the results, the connections seem logical. We can see that in the same household characters are normally allies. For instance Jon and Robb Stark, or Arya and Sansa Stark. There are also good relations between Joffrey and Sansa, which were going to marry. We can see that the algorithm can also be faulty at times. Sentiment analysis didnt work that well in some relations such as Jon and Ned Stark. The connection is very dark, but that is probably due to the fact, that Jon was often referenced as Ned Starks bastard, which can affect sentiment analysis.

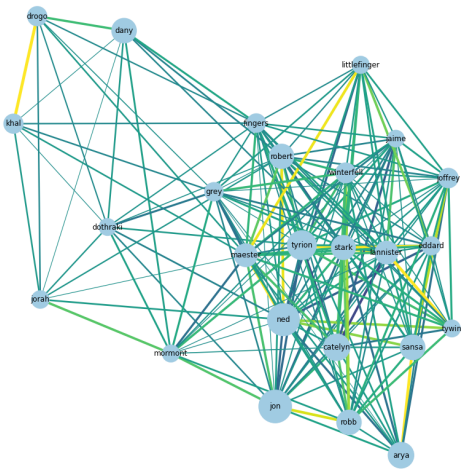

Figure 1. Sentiment graph of characters in Game of Thrones

Next steps

We can still improve our preprocessing to better tailor it for the Game of Thrones books and remove words that are recognised as entities.

We will also try using different algorithms for named entity recognition for instance BERT, or machine learning-based models and improve entity extraction.

We can improve NER by also predetermining names of all the characters with web scraping and then use these names to find people in the book. With the help of this prior knowledge, we will make better filters and thus get better nodes. This will allow us to increase the number of nodes shown and and because of this we will be able to focus on better network analysis (e.g. search for alliances).

We could also try to incorporate stemming to replace all nicknames and abreviations of character with their actual names to avoid recognising characters multiple times.

We can finetune relationship extraction to determine a class between each pair of nodes (eg. ally and enemy) so that there will be a clearer differentiation between households.

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.