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.

Related Work

J. Leng and P. Jiang explored relationship extraction using a deep learning approach to help support decision making in a social manufacturing context. This paper focused on creating a deep learning model that is based on an improved stacked denoising auto-encoder on sentence-level features. This model extracts manufacturing relationships, named entities, and text-based contexts. Their results reveal that this approach produces similar performance to the latest learning models. [1]

C. Giles and J. Wren developed an SVM classifier using NLP-based approaches to extract relationships between entities such as genes, chemicals, metabolites, phenotypes, and diseases. The classifier would label what kind of relationship there is between two entities, such as positive or negative, then detect what type of semantic relationship they have, such as binds, cleaves, etc. The main goal of this paper was to get the highest accuracy possible to prevent contradictions, but some inaccuracy was expected. SVM cross-validation showed that the implementation had a good F1 score at 88%. [2]

A. Roberts et al. worked on a supervised machine learning system (SVM) which was trained on a corpus of patient narratives which were manually annotated with clinical relationships. To evaluate the model, a standard ten-fold cross-validation methodology and standard evaluation metrics were used. This resulted in an F1 score of 75%. The paper concluded that it is possible to extract (clinical) relationships from text using an SVM with shallow features. [3]

V. Devisree and PC. Reghu Raj combined the unsupervised and supervised approaches to create a hybrid model that is able to extract relationships from stories. The model identifies the main characters, collects sentences that relate to them, then these sentences are analysed, and finally, the relationship between the characters is classified. Whilst training, if the sentence is classified, then it goes through the supervised section of the model, and if it's unclassified, it goes through the unsupervised section of the model.1 The corpus used was a set of 100 short stories for kids, which contained multiple different relationship types. Three relationship categories were used; Parent-Child, Friendship, and No-Relation, the hybrid model managed to achieve an F1 score of 83% [4].

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. We took two datasets of stories, both based on data from Project Gutenberg. We used NLTK to split each story into sentences and selected the stories with at most 1 000 sentences. Then we used spaCy's English language model to tokenise the sentences and search each story for words indicating family relationships. We calculated the average number of such words per sentence and selected only the stories where at least 5% of sentences contain such words.

This resulted in a total of 96 candidate stories; however, not all of them were suitable as they didn't contain many named characters. Such stories might, for example, contain references to the mother or father of the main character but never state their names. That's why we took spaCy's transformer-based English language model to do NER (named entity recognition) and filter out the stories with less than 4 named persons.

In the end, we got 47 short stories and a rough list of recognised persons for each story. The goal was first to improve the NER model to fix incorrect labels and detect the "poetic names," e.g. when a person is referred to as "the Bearded One" throughout the story. Then we needed to group all names belonging to the same person, and add the data about family relationships, so it would be suitable for training DocRed [5]-derived models for document-level relation extraction, such as DocuNet [6].

Data preparation

We used Prodigy to improve data annotation, as it's aimed to complement spaCy, and it offers a really simple UI as well as on-line model training for semi-supervised data annotation. The resulting model can then directly be used with spaCy on new data.

Stories generally don't explicitly state relationships between persons in a single sentence; it's much more common for the relationships to be evident from several indirect sentences. That's why it's necessary to label such indirect sentences as "evidence" of a relationship between two persons. That's why the final dataset we prepared contains tokenised sentences, named entities (persons) and all different names each entity is referred to in the text, relationships between those entities, and finally, which sentences serve as evidence of such relationships.

Results

The resulting NER model achieved accuracy above 90% on the stories, which is a significant improvement over 67% achieved by spaCy's RoBERTa-based model for the English language.

The final relation extraction model, however, wasn't able to achieve the expected results, as it quickly started overfitting due to a relatively small dataset and only achieved the best F1 score of 4%.

Discussion

Clearly, a dataset of 47 stories isn't sufficient to train a large NLP model, and significantly more data would be needed to achieve a sufficient improvement or come close to the state-of-the-art models for document-level relation extraction of over 60%. To speed up data annotation for such a dataset, it would make sense to define a custom config for Prodigy where all labelling could be done in a single step; however, for our purposes and the scale of the problem, it would take longer to do that than it did to do the last step of labelling manually in a text editor.

References

[1] Jiewu Leng and Pingyu Jiang. A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm. *Knowledge-Based Systems*, 100:188–199, 2016.

[2] Cory B Giles and Jonathan D Wren. Large-scale directional relationship extraction and resolution. In *BMC bioinformatics*, volume 9, pages 1–13. BioMed Central, 2008.

[3] Angus Roberts, Robert Gaizauskas, and Mark Hepple. Extracting clinical relationships from patient narratives. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 10–18, 2008.

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

[5] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of ACL 2019*, 2019.

[6] 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.