Relation Extraction on FoodDisease Dataset
-TUW-
Natural Language Processing

Ganellari Artjola - Godun Alina - Habenicht Richard - Kiassa Ouassim

1. INTRODUCTION

In the recent years, relation extraction has found empty spaces of possible implementations in various fields, delivering values and improvements to domains such as medicine, everyday life of human interactions etc,. Relation extraction is a crucial development because it allows for the automatic extraction of structured information from unstructured text, which can be used for a variety of tasks such as information retrieval, question answering, and knowledge base construction. This can improve the efficiency and accuracy of these tasks, and enable new applications that rely on extracting structured information from text.

In this project, we are going to present various classification methods that learn to predict relationships between entities based on the medical information given by the Food Disease Dataset [1].

1.1. Problem statement

With the increasing amount of data, especially in the medical domain, people are continuously in need of automatized solutions to understand the data properly, and make use of these solutions to structure the information and understand relations retrieved from it. In the medical domain, this remains a challengind task because of the complexity and technical nature of the language used, as well as the lack of standardization in terms and phrases used to describe entities and relationships.

Considering the above statements, in this project we are going to develop and implement baseline and deep learning models to accurately recognize the presence of cause, treat and neutral relations from the Food Disease dataset that will be presented in the section below.

1.2 Dataset

The main dataset used for this project is the Food Disease Dataset [1] which contains 608 medical sentences from abstracts of PubMed articles, with cause and treat relations between food and disease entities. This is a quite a small and imbalanced dataset according to the appearance in the dataset of each of the classes (323 treat, 144 neutral, 141 cause labels). The variables from this dataset that will be used for modeling will be only the sentence (containing medical information on food and diseases) and label (treat, cause, and neutral).

As a supplementary dataset, we have also used the Crowdtruth dataset [2], which contains annotated data for cause and treat relations in sentences from abstracts of PubMed articles. The dataset contains 4028 sentences annotated for the existence of a cause relation and 3983 sentences annotated for the existence of a treat relation. Again we have used only the sentence and label variable (treat, cause).

2. IMPLEMENTED SOLUTION

In order to be more concise, we are going to separate the implemented solutions into three stages, initial (milestone 1), intermediate (milestone 2) and final stage (milestone 3).

In the first milestone we have implemented some baseline models on the Food disease dataset such as Naive Bayes from the nltk package, Multinomial Naive Bayes and SVM from sklearn. After some preprocessing steps such as stopwords removal, lemmatization on the words of the sentences and transformation into a dictionary of the data, these models have been trained and then tested using first the three classes, then only two classes (treat and cause) and finally undersampling the overrepresented treat class because of the preference that the model showed to this class. For simplification purposes, below in table 1 we will present the performance metrics only for the highest performing model with was SVM and binary SVM. The classification result for each of the methods used in presented in the table 1 below:

	Cause		Treat		Neutral	
Model	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**
SVM	0.74	0.69	0.72	0.91	0.83	0.54
SVM binary	0.82	0.72	0.90	0.94	-	-

Table 1. Milestone I performance metrics

	Cause		Treat		Neutral	
Model	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**
Bi-LSTM	0.85	0.38	0.72	0.97	0.81	0.61
Bi-RNN ML	0.90	0.32	0.62	1.00	1.00	0.24

Table 2. Milestone II performance metrics

In the second milestone, we implemented some different approaches, including DNN algorithms such as NN, LSTM, and baseline NN model. These highest-performing models are presented in table 2.

Finally, in the third milestone we implemented and tested various methods such as using bigrams and trigrams on SGD classifier, LSTM model trained on Crowdtruth dataset and tested on Food disease (transfer learning), experimenting with binary classification on models including neutral vs non_neutral and cause vs treat, and part of speech tagging (POS). From the conducted experiments, the method that improved the models from previous milestones was feeding a binary classification with bigrams as an input. The recall increased from 0.72 to 0.97 for the treat class.

3. CHALLENGES

The main challenge to this task was the fact that we had to experiment with an imbalanced and small dataset, and even though we explored the feasibility of different baseline and DNN models, n-gram language models, binary classification, POS etc, we still could not fully ignore its presence, which was shown on the performance metrics that we acquired. Very frequently the neural network approaches all had a tendency for overfitting on such a small dataset and that was also one of the reasons why we tested bigrams/trigrams on one of the "simpler" models from milestone 1.

4. EXTERNAL RESOURCES

All the experiments have been conducted using the Python language and ML framework of Pytorch to build the DNN models, trained and tested on the aforementioned dataset, Crowdtruth and FoodDisease.

5. DISCUSSION

Implementing ML models on an unbalanced and small medical dataset is always a challenge that requires special attention to finding methods that could possibly overcome it. The methods and models implemented during this project, have suggested that the abovementioned reason, is the main cause to not obtaining a good performance. For example, during the second milestone, we concluded that many words from the test dataset of Food Disease never occur in the train dataset. That is we considered including the Crowdtruth dataset to increase the dataset and vocabulary size. Using transfer learning on the bi-LSTM model on both Crowdtruth and FoodDisease, did not improve the classification accuracy because of the food disease dataset having specific sentences, while the Crowdtruth dataset has more sentences but only a small percentage of these sentences have any relation with food entities. Furthermore, we utilized undersampling techniques in some of the models such as Bi-LSTM, Multi NB, and SVM but it still did not improve the performance metrics. With all the experiments conducted, the model that performed better in comparison to the others is the SGD classifier trained on bigrams as inputs only on two classes (treat and cause), with a recall of 0.97 for treat and 0.8 for the cause class.

In conclusion, our current work presents some experiments on extracting relations between food and disease in FoodDisease dataset, but there is still room for improvement. One potential direction would be to use a transformer like BioBERT to gain a much larger vocabulary and fine-tune this transformer for both Crowdtruth and FoodDisease datasets. Additionally, dependency parsing and position indicators to analyze the grammatical structure of the sentences and find out related words as well as the type of their relationship between them might help in achieving a higher classification accuracy. Finally, including knowledge or dependency graphs by using the Potato package, might show improvements and would add more explainability to the models.

References

[1] Gjorgjina Cenikj, Tome Eftimov, and Barbara Korouˇsi´c Seljak. “SAFFRON: tranSfer leArning For Food-disease RelatiOn extractioN”. In: Proceedings of the 20th Workshop on Biomedical Language Processing. Online: Association for Computational Linguistics, June 2021, pp. 30–40. doi: 10.18653/v1/2021.bionlp-1.4. url: https://aclanthology.org/2021. Bionlp-1.4.

[2] Anca Dumitrache, Lora Aroyo, and Chris Welty. “Crowdsourcing Ground Truth for Medical Relation Extraction”. In: ACM Transactions on Interactive Intelligent Systems 8.2 (2018), 1–20. issn: 2160-6463. doi: 10 . 1145 / 3152889. url: http : / / dx . doi . org / 10 . 1145 / 3152889