A Semantic Relation Extraction Model Application

Aleksandr Sergeev
Earlham College

Background

In recent years we have seen a growth in the amount of datasets and models built and trained on them to recognize patterns in unstructured text. Many contributors are creating state-of-the-art machine learning models capable of identifying semantic relations in these texts, organizing information into categories.

With increasing amounts of training data available, contributors are raising the levels of accuracy their models can achieve, which creates higher applied orderings to the unstructured text, making it easier to identify sought-after knowledge about subjects of interest.

Project Description

This project aims to build a machine learning model capable of labelling sentences outside of its training data to observe the accuracy on foreign data. Successful classification of this foreign material would decrease the amount of extra text a user would have to go through to find what they want, increasing speeds of information retreval and making this a tool for information extraction.

Methodology

Figure 1 below shows the architecture of this software. A user submits data in .pdf or .txt form, where it will have text extracted and tokenized, followed by entering entities of interest. Sentences containing the queried entity are preprocessed and put through the Keras model. The user may make multiple queries and save classifications from their requests to disk.

Training new models and validating their accuracy. This makes the project more open so that the user may experiment with different classification techniques and model implementations.

This project uses the following machine learning libraries running on Python3.6; Keras, Scikit-learn, Spacy.

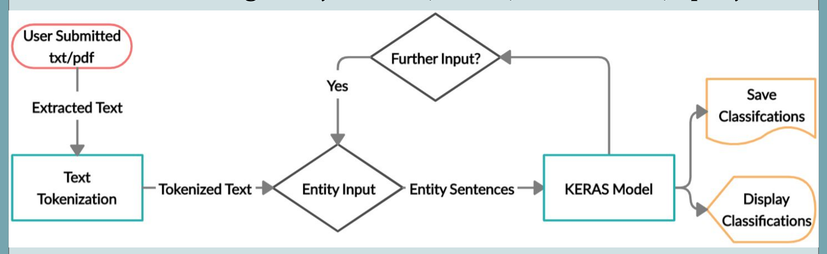

Figure 1: Software Architecture

The Keras LSTM classifier built and trained for this project used a Word2Vec pretrained embedding layer, followed by the Bidirectional LSTM layer for sequence classification and two dense layers to reach the final output. It was trained on the SemEval 2010 Task 8 dataset for semantic relation classification. The embedding layer did not allow for further training of parameters.

Results

The BI-LSTM classifier was trained for 20 epochs, with Figure 2 showing the progression of training accuracy and loss through each epoch. This model achieved a 96.2% validation accuracy after training on the whole dataset, and 78% on the test batch at the end of training.

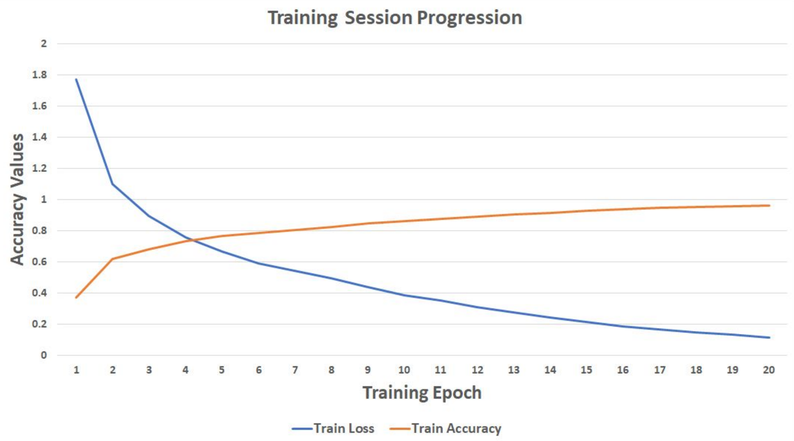

Figure 2: Line graph showing test accuracy and loss over training epochs

Valid sentences for processing require two tagged entities. The dataset provided a total of ten possible labels.

The 'other' label was the most commonly predicted, shown in figure 3 to be 33% on average. This was checked using four very different source materials. The model also had other classes where the percentage of classifications were similar. There was therefore a pattern in the distribution of label predictions over a variety of material.

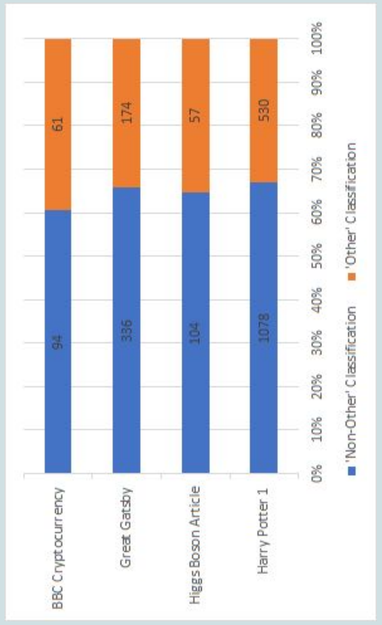

Figure 3: Percentages of Classifications

Discussion

The model performed well after some epochs on the dataset, but working with foreign material presented flaws. The material used was fundamentally different; education articles, different genre novels. Noticeable prediction patterns did occur in the results which indicates the model was filling quotas of predictions.

The BI-LSTM model used was not very complex and lacked detailed attention to the task. It is likely that State-of-the-art models would have better predictions when configured to work with foreign data.

Figure 4 shows the confusion matrix, indicating that some classes had more examples, while others lacked enough, which may explain results from Figure 3.

	Cause-Effect	Component-Whole	Content-Container	Entity-Destination	Entity-Origin	Instrument-Agency	Member-Collection	Message-Topic	Other	Product-Producer
Cause-Effect	1294	0	0	0	6	1	0	4	6	4
Component-Whole	0	1190	4	1	3	8	6	13	20	1
Content-Container	0	3	721	1	0	0	0	0	3	0
Entity-Destination	0	0	21	1093	0	0	0	1	10	0
Entity-Origin	5	1	0	1	941	2	0	1	9	1
Instrument-Agency	0	3	1	0	0	623	1	0	17	7
Member-Collection	0	3	0	1	0	1	889	1	16	1
Message-Topic	0	1	0	0	2	0	0	872	13	0
Other	16	23	11	24	28	10	19	15	1675	23
Product-Producer	4	2	0	1	7	3	0	1	12	909

Figure 4: Confusion Matrix on Dataset

Future Work

In the recent years, most sophisticated and larger datasets have been released with more examples and more classes of information. Increasing the size of the dataset used would provide the model with better chance at identify key characteristics of certain labels to help it predict outside of a dataset. Refining of the model to the task by introducing more layers (Attention layer is common) would also improve the work.

Acknowledgements

I would like to thank Professor Charlie Peck for his mentorship over the years and help in finding this idea. Also to Professor David Barbella as my advisor for his valuable feedback and guidance throughout the development of the project. Finally, a thank you to the department as a whole for their openness to support me and the rest of my peers always.