Natural Language Processing

Group 15

Report

01. Introduction

The aim of this project was a typical relation extraction problem: taking sentences as input and identifying the relational connection between the two most prominent entities within them. To do that, the team decided to use a domain-independent dataset called "SemEval-2010 Task 8"[1] to evaluate different approaches and identify the strengths and weaknesses of them. The task was especially challenging as the dataset was not only rather small (8,000 sentences to train on), but also had 10 target classes to predict.

The project was structured in the following way: After deciding on and implementing a baseline classifier for the first milestone to get a feeling for the expected performance of a model in the problem space, the team then selected a deep learning model for the second milestone to further enhance the performance and get more insights into the expected results. On top, analyzing the results, identifying the relevant factors for a successful prediction and aiming to improve on them formed the final step of the project.

02. Approach and Results

As mentioned in the introduction, the project started off with a baseline model where the team after careful consideration and research decided to apply a Multinomial Naive Bayes classifier from the scikit-learn library as this was an efficient model being proposed multiple times in similar projects. For preprocessing a CountVectorizer was used to tokenize the text and a TfidfTransformer retrieved the term and inverse document frequencies of the input sentences. The result was an accuracy score of 60.5% (F1-Score: 0.78) - this formed our first baseline, although later analysis will go into more detail.

For the second milestone, a Bag-of-Words (BoW) classifier was used to represent a deep learning approach - which had difficulties with the dataset in its main setup and performed rather poorly with an F1-Score of 58.2%. Instead of trying to tune model (hyper-)parameters in order to improve the results, the team looked deeper into the data in order to get a better feeling of the actual problem at hand.

After very important and helpful feedback from the team's advisor Gábor Recski, instead of trying to solve the entire problem as a multiclass problem, to ensure a better understanding of the data it was modified to a binary classification problem: instead of asking "In which of these 10 categories does that relationship fall?" we ask "Is this relationship a [relationship] or not (=other)?". This modification opened up space for a deeper analysis and for forming hypotheses and assumptions about the models' logics and mechanics that fundamentally influence the prediction process.

When focusing on the relationship "Entity-Destination" and its correct and wrong predictions of the Multinomial Naive Bayes classifier, it seems as if the model does a relatively straightforward mapping of

[1] https://aclanthology.org/S10-1006/

keywords it categorizes as relevant for the relationship such as “into” or “to” which appear disproportionately often in all predictions where the model thinks it’s the actual relation. Unfortunately, the English language is more complex in that regard as there are many instances were these keywords are used in another, sometimes a more abstract, context (“A daughter donated her kidney to her father.”), which makes some examples even difficult for members of the team to accurately assign to the correct class (“A man threw a child into an outlet.”). This once again underlines the difficulty of the task in combination with the dataset.

In a second round the team aimed for helping the model to better understand the different meanings and context situations of the same words (“into” can mean different things in different situations and does not necessarily introduce a destination) by applying part of speech (PoS) as a concept of preprocessing the text before the training phase. The word-tokens are enhanced with meta information (e.g. whether a word is a noun or an adjective) which in theory should help the model to perform better - unfortunately, the results did not change significantly.

03. Learnings

Apart from many technical tools that we have become accustomed to as part of the course such as tokenization, other preprocessing techniques as well as many different (scikit-learn) models, we have learned and experienced firsthand the challenges of modeling, learning from and predicting natural (English) language in a classification setting. Applying methods being taught in university courses is one thing, trying to go the extra mile and uncovering the reasons for the performance of (partly rather intransparent) models is another thing. The main challenge for us as a team was the problem-analysis as we were lacking frameworks to do so and were oftenly discussing ideas and assumptions - without finding efficient ways to properly check them in practice in the code. Choosing a different data set or data domain respectively in general might have helped although difficult to tell.

04. Discussion

Several baseline classifiers ranging from simple scikit-learn implementations to a little bit more complex deep learning models with rather mediocre performance results form the final status of the project. A deep dive into the results and their patterns after the simplification of the dataset to a binary problem improved the team’s understanding of how the models try to find logic in the small amount of data, however a much more sophisticated approach and/or bigger dataset might enable better prediction results as even the simpler binary problem did not deliver results we as a team would be proud to report.