Text classifications using IMapBook dataset

Domen Kos, Timotej Kovač
University of Ljubljana
Faculty of computer and information science
Večna pot 113, SI-1000 Ljubljana
dk6314@student.uni-lj.si, tk3713@student.uni-lj.si

Abstract

TODO

1 Introduction

For the course assignment at *Natural language processing* class we decided to do text classification on IMapBook dataset. The dataset contains short discussions between primary school students who are chatting about different book topics. Each record is annotated with 16 attributes. Original messages are posted in Slovene language, but they are also translated to English. We also have the information about topic they are discussing, if their message was an answer to some previous asked question and if their discussion is relevant to the topic, since there are no constraints so they can write anything they want. If the discussion is moving away from the proposed topic, the teacher can intervene and guides it back by asking some questions relevant to the book. Dataset contains approximately 3500 messages about 3 different short stories. Our goal is to develop the models which could detect the topic of the current debate so the teacher can intervene. The idea is to develop different models and combine their outputs. We would analyse conversation on different levels. First would be to analyse separate messages and define their relevance to the topic. We would also define type of message which can either be answer or question and also the category. By combining results of separate messages we could determine when the conversation is starting to move away from the topic.

2 Related work

Text classification is a popular topic of natural language processing (NLP) field, thus there are many researchers working on the field. In this section we present most relevant work and techniques they used.

Most of the research work and the state-of-the-art results were achieved using English language, but some of the techniques and approaches can be adapted to Slovene language. The authors of paper (Fernández Anta et al., 2013) had similar problem, where they analysed Spanish tweets which are also relative short texts. They discuss different approaches for preprocessing data to extract the most relevant features which are later used for classification. Few standard approaches are discussed like how to define a basic term which are used by classification algorithms. Such as uni-grams (1-grams), bi-grams (2-grams), tri-grams (3-grams), n-grams. They found out that having n larger than 3 does not improve results. They also tried combining different types of n-grams (like uni-grams and bi-grams) to achieve better results. That also means that attribute list was larger so they removed some entries by setting a threshold value. With threshold they removed n-grams that did not appear frequently enough or they appeared to many times, since they were considered as noise. They also discuss how important are stemming and lemmatization comparing English and Spanish language which is also interesting for Slovene language which is also morphological richer then English.

In second paper (Lai et al., 2015) authors presented the recurrent convolutional neural network for text classification. The network was able to capture contextual information of the sentence and extract features and learn some word representation. As described the disadvantage of the recurrent network is that although the context of the word is captured, the model is biased where later words are more dominant then earlier. To tackle this problem they applied additional convolutional and pooling layers. They learned a word Representation and used it for some text classification.

Another similar approach is used in third paper (Johnson and Zhang, 2014) where the authors

used convolutional neural networks for text categorization where the word order was also taken into account. The input to the network is not standard bag-of-word representation but they present their own word representations which are some higher dimensional vectors where 2D convolution is applied. For baseline model they used a support vector machine (SVM) classifier with bag-of-word representation and showed that their approach gave them lower error rate than standard models.

There are also many already pre-trained word representations which can be used, also for Slovenian language. The word embeddings induced from a large collection of Slovene texts composed of existing corpora of Slovene were prepared and published on CLARIN (Ljubešić and Erjavec, 2018). This could also be useful with our task since the embeddings were learned on bigger corpuses then are available to us.

3 Initial ideas

To determine if the teacher must intervene we need to answer the following questions:

- are the messages book relevant,
- what type is the message,
- in what category does it belong.

Based on this information we could then determine if the conversation is in need of an intervention or not.

Because there are three seperate requirements our first idea was to come up with three seperate classifiers. We will start with standard text classification procedures like tokenizing, stemming, removal of stop words and then represented words as vectors in order to use them in our machine learning algorithms. After that we will probably use some kind of machine-learning approach. Recurring convolutional neural networks or SVMs could be used to make use of the sequential information of words as well.

3.1 Book relevance

Here the answer we are trying to answer is whether the message is related to a story or not. From the data itself came to some conclusions:

- Category of the message is a good indication whether the message is book relevant. So if the message is classified as having a category discussion it is a good chance that the message is book relevant. So the result of the message category classifier could be used here to determine if the book is relevant.
- Conversations have some retention. If the conversation starts leaning towards a discussion of a book most messages will be about the book, and if the conversation starts to move towards some other category most of the messages will follow. So here the sequence and previous states could be deemed important.

So maybe some of these observation might help us especcialy if we used the results of some of the other ones. We will try to experiment with how to include this additional knowledge and if it brings any improvements.

3.2 Type of the message

Here we try to answer the type of the message. This can be a statement, a question or an answer. Here too we drew some conclusions from the data available:

- Answers tend to follow questions.
- Answers are mostly regarded as book relevant and statements are not.

These conclusions might help us to came up with a better model.

3.3 Message category

Each message can be one of the following:

- chatting,
- switching,
- discussion,
- moderating,
- identity or
- other.

When determening the message category we will try to include the order of the messages as well.

References

Antonio Fernández Anta, Philippe Morere, Luis Chiroque, and A. Santos. 2013. Sentiment analysis and topic detection of spanish tweets: A comparative study of nlp techniques. *Procesamiento de Lenguaje Natural*, 50:45–52.

Rie Johnson and Tong Zhang. 2014. Effective use of word order for text categorization with convolutional neural networks. *CoRR*, abs/1412.1058.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Nikola Ljubešić and Tomaž Erjavec. 2018. Word embeddings CLARIN.SI-embed.sl 1.0. Slovenian language resource repository CLARIN.SI.