SI 630 - Winter 2020
Reading Response 1

Yirui Gao

January 21, 2020

Overall summary

This paper majorly discusses the findings of tokenisation strategies and classification models that achieve to train a good-performanced language identification model. The authors expanded the paper by firstly talking about the background of language identification and the difficulty for doing this under the growing content for web information. With the known classification models like Nearest Neighbors, Naive Bayes and SVM, and known tokenisation methods, like byte tokenisation and codepoint tokenisation, they used the combinations of different sets of methods combining different ways of measurement of similarity to search for the best model for the main task. They utilized three datasets with different properties, like the number of encodings, the length and total number of the documents and the number of languages, to additionally dig out the influence of these aspects on the performance of the language identification model.

They concluded by showing the experimental results that models that outperform others generally are a simple 1-NN model with cosine similarity and an SVM with a linear kernel. And patterns can be found that the language identification task might tend to be easy to deal with small number of languages with long documents, but much harder to deal with large numbers of languages, shorted documents and greater class skew. These results tend to verify the presumptions that if documents have multiple languages, the difficulty of the identification task must be large. This study is quite important and since it was published in 2010, the results that it provided great helpful guidance for language identification tasks later by pointing out the models that perform well.

Main contributions

I think the main contributions for this paper, are first, they verified the difficulty of language identification task vary under different given conditions like the number of languages, the length of the document, etc. This is quite important since it gives an overall basis and guides this task with experimental results. Second, they found that among various tokenisation methods, the byte-based tokenisation without character encoding detection is better than the codepoint-based tokenisation with character encoding detection. Also, a simple cosine similarity-based Nearest Neighbor is a better classifier than others like SVM and Naive Bayes. These findings also offered great help for later research since it technically pointed out the methods that should firstly be considered when working on this task.

Fill Out Three of the Following Responses (Your Choice)

3.1 What would I have done differently?

There are majorly two concerns for me about this paper. The first one is about the methodology of this experiement, especially about the usage of classifier. My concern is, since they mentioned in the introduction that they would like to find out the best combination of tokenisation strategies and classification model, but they only used three classification models, including NN, NB and SVM. I doubt is it enough to carry out comparative experiments. If possible, I think other popular classification methods like Logistic regression and decision tree should also be applied and compared.

The second concern is that is it possible if they could conduct other set of comparative studies to dig out which property brings the largest difficulty to the language identification task. Since they proposed and tested that the number of languages, document lengths and encoding all influence the performance, but they did not mention which one dominates the influence, and maybe they could use experiments to verify that which factor should we care the most and for which, we don't have to focus a lot on it.

3.2 What was I confused about?

I still have several confusions after finishing the reading for this paper:

- How are byte-based tokenisation and codepoint-based tokenisation working differently? This question is a fundamental one but I think due to it is a basic background knowledge, the paper doesn't stress the difference of these two in detail.
- When mentioning about the results for differing n-gram size, the authors mentioned that for Skew with 1-NN model, "rather than using a constant α value for all n-gram orders, it may be better to parameterise it using an exponential scale use such as $\alpha = 1 - \beta^n$ ". How this conclusion results from?

3.3 Questions for the author(s)

Since this is paper published ten years ago, and we know, starting from the year 2014, great use of neural net benefits all fields in the machine learning and NLP fields. If I could have a chance with the authors, I would like to know whether using the new deep learning methods, like combinations of CNN and LSTM, or other things, will the results become different? Maybe the authors have done the new job based on deep learning, but for this paper, are the methods used in this paper still up to date, or already out of date? These questions reflect that with the rapid development of artificial intelligence, some previous research outcomes might tend to be not that useful for the current researches. We need to question ourselves all the time: whether this method stated here still works well for my task?

References

Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, USA, HLT '10, page 229–237.