Proposal

Jeremy Lachowicz

February 11, 2022

1 Problem

In this project I will attempt to discover which word embedding technique (Word2Vec or Doc2Vec) is better for authorship attribution in the Ancient Greek language with the hopes of applying this to determine the author of Hebrews.

2 Literature Survey

In this paper, [ACHII18] , they used different word embedding tehcniques for authorship attribution in the Bengali language. They had a corpus of roughly 2,250,000 words per author, with six authors. They found Word2Vec to work better with skip-grams than continuous bag of words. They used three different types of deep neural networks to train on the feature sets: ANN, RNN, and CNN. They found the CNN model to perform the best on Word2Vec with skip-grams and CBOW's.

In this paper, [BAH+19], they wanted to see if Word2Vec is a good method of authorship attribution. They had a corpus of 72 documents with 8 authors (9 documents each author), although they did not state the length of the documents. They found Word2Vec was very accurate with authorship attribution and CBOW produced better results than skip-grams. They also explained which hyper parameters they used for the Word2Vec model: LayerSize - 50, WindowSize - 5, MinWordFrequency - 1, Iterations - 3, LearningRate - 1.0E-4, Sampling - 1.0E-5.

In this paper, [GAPDSP18], they wanted to figure out which Doc2Vec hyperparamters and feature types are best for Doc2Vec authorship attribution. They had a corpus consisting of about 30 articles per author from 1999 to 2009 from different topics in the *The Guardian* newspaper. The average document contained about 1,000 words. For feature sets they used n-grams (ranging from 1-5), POS tags, and words, with a logistic regression classifier. They found using a combination of n-grams as a feature set was the best.

In this paper, [RBGR+20], they wanted to compare different models for authorship attribution and use an ensemble to see if that creates better results. They used a corpus of English, Italian, Spanish, and French texts, but they did not state the length or number of documents. They tested Word2Vec and Doc2Vec and found Word2Vec was better overall, as well using a logisitc regressor for their ensemble classifier. Then, the feature sets they used were N-grams, Word2Vec, and TF-IDF. They found Word2Vec to perform the best out of the three, but when combined in an ensemble, it sometimes performed even better.

3 Methodology

First, I will compile a second set of Ancient Greek texts with similar sizes to test Doc2Vec and Word2Vec to discover which produces the best classifier for Ancient Greek.

For preprocessing the data, previous research indicates tokenizing the text and then removing stop words, stemming words, and removing punctuation is the standard for authorship attribution. I was unable to find a stemmer for Ancient Greek, but there is a tokenizer and stop-word remover for Ancient Greek in the CLTK (Classical Language Toolkit) library. Therefore, I will plan to tokenize, remove stop-words, remove punctuation, and remove any numbers in the text that indicate chapter or line.

For the model of Word2Vec, the options are CBOW's (continuous bag of words) or skip-grams. There is conflicting research on whether to use skip-grams or CBOW, so I will test both of them. The

following hyperparameters were used with success in another paper: LayerSize - 50, WindowSize - 5, MinWordFrequency - 1, Iterations - 3, LearningRate - 1.0E-4, Sampling - 1.0E-5. However, due to the differences in language, I will try multiple hyperparameters but use these recommended parameters as my "middle" numbers in the grid search. From here I will decided which Word2Vec model is better based on accuracy of determining authorship of the known texts.

For the model of Doc2Vec, the options are DBOW (distributed bag of words) and DM (distributed memory). I will test both of these as one is not outright better than the other for authorship attribution. For hyperparamters I will run a grid search to find the best. For the feature sets, previous research has indicated it is best to use a combination of n-grams on the words in the text (rather than parts-of-speech tags or characters), specifically combining 1-grams and 2-grams by concatenating their respective models (vectors). Therefore, I will use this combination of n-grams. From here I will decided which Word2Vec model is better based on accuracy of determining authorship of the known texts. Also, to train this I will consider individual paragraphs as a document.

For both the Word2Vec and Doc2Vec, I will use a Logistic Regression classifier to determine authorship. From prior research, its seems a logistic regression model is considered the best for authorship attribution, used with default parameters. Finally, I will choose the better model and test it on the Hebrews data.

If I have time, I want to combine DBOW and DM for Doc2Vec to see if that gives me better results. I also would like to try an ensemble model and potentially combine Word2Vec and Doc2Vec with other higher performing models.

4 Timeline

- Week 5: Code the Word2Vec process for the first set of Greek sources.
- Week 6: Finish coding the Word2Vec process for the first set of Greek sources.
- Week 7: Code the Doc2Vec process for the first set of Greek sources.
- Week 8: Finish coding the Doc2Vec process for the first set of Greek sources.
- Week 9: Test each technique and decide which one is better.
- Week 10: Use the best technique on the Hebrews sources.
- Week 11: Write final report.
- Week 12: Write final report and present.

References

[ACHII18] Hemayet Ahmed Chowdhury, Md. Azizul Haque Imon, and Md. Saiful Islam. A comparative analysis of word embedding representations in authorship attribution of bengali literature. In *2018 21st International Conference of Computer and Information Technology (ICCIT)*, page 1–6. IEEE, Dec 2018.

[BAH+19] Nacer Eddine Benzebouchi, Nabiha Azizi, Nacer Eddine Hammami, Didier Schwab, Mohammed Chiheb Eddine Khelaifia, and Monther Aldwairi. Authors' writing styles based authorship identification system using the text representation vector. In *2019 16th International Multi-Conference on Systems, Signals 'I&' Devices (SSD)*, page 371–376. IEEE, Mar 2019.

[GAPDSP18] Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Grigori Sidorov, and David Pinto. Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*, 100(7):741–756, Jul 2018.

[RBGR+20] Mostafa Rahgouy, Hamed Babaei Giglou, Taher Rahgooy, Mohammad Karami, and Erfan Mohammadzadeh. Cross-domain authorship attribution: Author identification using a multi-aspect ensemble approach notebook for pan at clef 2019. 07 2020.