Proposal

Jeremy Lachowicz

February 26, 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

Corpus Size and Language:

For Word2Vec and Doc2Vec having a large Corpora is important. [ACHII18] used roughly 2,250,000 words per author. Another paper used 72 documents per author [BAH+19], while another used 30 [GAPDSP18]. Word2Vec and Doc2Vec have been tested in other languages, namely Bengali [ACHII18] and English, Italian, Spanish, and French [RBGR+20]

Corpus Preprocessing:

For Word2Vec there are two options for input, continuous bag of words and skip-grams [ACHII18]. Prior research has found conflicting results for which technique is better. [ACHII18] found skip-grams to work better and [BAH+19] found CBOW to work better. For Doc2Vec there are two options for input, the options are distributed bag of words and distributed memory. They did not find a significant difference between the two. [GAPDSP18]

Hyperparameters and Feature Sets:

The Word2Vec hyperparamters that received significant results were: LayerSize - 50, WindowSize - 5, MinWordFrequency - 1, Iterations - 3, LearningRate - 1.0E-4, Sampling - 1.0E-5 [BAH+19]. One paper found combining n-grams (ranging from 1-5) returned the best results overall [GAPDSP18].

Other Models Used

[RBGR+20] used an ensemble of TF-IDF, N-grams, and Word2Vec. They found the results to sometimes be better in an ensemble, but not always. Another paper, [ACHII18] used three types of neural networks (ANN, RNN, and CNN). They found CNN model to perform the best with Word2Vec.

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 which model produces the most similar vector to the test set's vector.

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 decide 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 extract a vector from Word2Vec to represent each author. From here I will use cosine similarity to test if the vectors are pointing in similar directions in the n-dimensional space. This similarity measure was chosen due because cosine similarity depends on the angle of the vector, rather than the magnitude. For high dimensional text data, angle is more important due to the varying word counts and vocabulary of the texts. I also plan to extract other features that could imply writing style such as frequency of word use. This will allow me to obtain multiple measures of which document is closest to the target document. 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 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.