ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #44939

Masked Language Modeling for Predicting Missing Words in Damaged Ancient Greek Texts

Mr. Kyle Riccardi
Danushka Bandara, Fairfield University

DANUSHKA BANDARA received the bachelor's degree in Electrical Engineering from the University of Moratuwa, Sri Lanka, in 2009. He received his master's and Ph.D. degrees in Computer Engineering and Electrical and Computer Engineering from Syracuse University, Syracuse, NY, USA, in 2013 and 2018, respectively. From 2019 to 2020, he worked as a Data Scientist at Corning Incorporated, Corning, NY, USA. Currently, he is an Assistant Professor of Computer Science and Engineering at Fairfield University, Fairfield, CT, USA. His Current research interests include Applied machine learning, Bioinformatics, Human-computer interaction, and Computational social science.

Masked Language Modeling for Predicting Missing Words in Damaged Ancient Greek Texts

Abstract:

The ancient Greek texts are valuable for understanding and learning about the history, culture, and nuances of ancient Greek life. The texts come in many forms, including papyri, fragments of pottery, etc. Due to the nature of these materials and degradation over time, some of these texts are missing words, and even entire phrases. This makes it difficult for historians to interpret the texts. The data for this project was collected from the Perseus Collection and the 1KGreek collection, which contains 250,000 unique sentences of ancient Greek literature. The dataset was preprocessed using the import classical language toolkit (CLTK) and sentences were normalized for better encodings. After the encoding was done all our data was split by sentences and then they were fed into a Distil Bert masked language model. The word piece tokenizer for this model was trained using a vocabulary list of 35,000 words. By using the DistilBert transformer model we were able to train a masked language model based on words to achieve a Hit@5 of 34 percent, Hit@10 of 35 percent, Hit@100 of 36 percent, and a perplexity of 1.04. This model can be a valuable aid for the historians' workflow in deciphering damaged ancient texts.

Introduction:

When understanding the life and culture of an ancient civilization, its texts can be an useful resource. More specifically, when talking about Ancient Greece many pieces of literature have been damaged/fragmented over time. This is a challenge for historians as they rely on educated guesses as to what the author may have meant in certain texts.

Assael et. al [1] did a deep dive into this topic using their groundbreaking work on ‘Ithaca’: a character-masked language model. Where their work is a great step into breaking down ancient Greek texts the goal of our project is to see if a word-level language model, using transformer [2] based models, can achieve decent results for this problem. Our goal is to achieve maximum performance using words to see how well the model can do when certain words are absent. This achievement would allow researchers the ability to further decode damaged texts even when a majority of the context is missing. In essence, if successful, we will be able to predict text in cases where parts of the whole sentence are missing.

Methodology:

Preprocessing:

The texts for this project were gathered from the Perseus and 1kgreek corpora. These texts once gathered were split by sentences. Once the process of splitting each text by sentence was complete. We further preprocessed the text. First, the text was cleared of accents and standardized to the lowercase letters using the classical language toolkit (CLTK) Python library [3]. Then the given sentences were filtered out for punctuation, as some texts had ample

punctuation, while others did not. Once all of this was done the CLTK library was used once more to filter out all English and Latin words within the Greek texts to make sure we had only Greek sentences. There is English and Latin footnote,subtitles, and contextual statements for understanding the basis of the texts. This process left us with 2.5 million Greek sentences.

Models:

After this was done, we created a word piece tokenizer the standard Bidirectional Encoder from Transformers (BERT) tokenizer [4]. The tokenizer was created using a vocabulary list of 35,000 words. Once this tokenizer was done training on our dataset it was then time for us to encode our 2.5 million examples using the tokenizer and split our dataset into a training set of 85 percent and a test set of 15 percent. Once this was done, we masked 15 percent of the training data to understand 1 word per sentence in the testing dataset. This decision was made to test whether the model can accurately determine missing words based on a single missing word. Once, the masking of the encoding was done it was fed through a DistilBERT model [5].

Experimentation:

Table 1: Paramters for running Disil BERT Model

Learning Rate	0.00005
Max Embedding Length	128
Epochs	10
Hidden Layer Size	768
Number of Attention Heads	12
Number of Hidden Layers	6

This model used the AdamW [6] as the optimizer and the model was developed using PyTorch [7]. The model was evaluated using Mean Reciprocal Rank (MRR) using the Hit@5 for the correct word appearing in the top 5 and Hit@10 meaning the correct word appears in the top 10. The model was also evaluated using perplexity to determine how well the model predicted examples.

Results:

Table 2: Results for Disil BERT Model

Hit@5	34 %
Hit@10	35%
perplexity	1.04

Discussion:

We found that the model is predicting the correct word in the top 5, 34 percent of the time, which is decent when considering the limited amount of sentences the model was trained on Greek BERT model [8], which was trained on modern day Greek was trained on 10 million articles. Based on scale our results was around 500MB and the Greek BERT model contained 29GB of data. The same can be said for our results at Hit@10.

Our results show that by only using words we can get a concrete understanding of the Ancient Greek language. This model also shows us that using the distilBERT framework alone can provide good word-level accuracy without the use of character-level modeling.

Conclusion:

In conclusion, this model shows good results, while also holding promise for future improvements. The next step for our research is to get feedback from historians on how useful the model is and how we can improve the output to be more useful to them. Limited data is a issue within multiple ancient languages due to texts being damaged through age, lost, or based on reliability of word of mouth. Another path for future research includes using such models in other ancient languages with limited data availability.

References:

Bibliography

(1)Assael, Y., Sommerschield, T., Shillingford, B. et al. Restoring and attributing ancient texts using deep neural networks. *Nature* **603**, 280–283 (2022). HYPERLINK "https://doi.org/10.1038/s41586-022-04448-z" https://doi.org/10.1038/s41586-022-04448-z

(2) Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, & Illia Polosukhin. (2023). Attention Is All You Need.

(3)Johnson, Kyle P., Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. "The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 20-29. 2021. 10.18653/v1/2021.acl-demo.3

(4)Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. HYPERLINK "https://aclanthology.org/N19-1423" BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding . In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

(5)Victor Sanh, Lysandre Debut, Julien Chaumond, & Thomas Wolf (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR, abs/1910.01108.*

(6) Ilya Loshchilov, & Frank Hutter. (2019). Decoupled Weight Decay Regularization.

(7) Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

(8) Koutsikakis, J., Chalkidis, I., Malakasiotis, P., & Androutsopoulos, I. (2020). GREEK-BERT: The Greeks visiting Sesame Street. In *11th Hellenic Conference on Artificial Intelligence*. ACM.