Effect of Context Vectors in the Task of Authorship Verification

Sai Kaushik Mudigonda, Master of Science in Computer Science

University of Dublin, Trinity College, 2022

Supervisor: Dr. Carl Vogel

Authorship Verification is a growing area in the field of natural language processing and text analysis. Due to the high availability of text online there is even more research into identifying and understanding the author. Other areas where this would be helpful would be plagiarism detection, citing proper authors, authorship profiling, obtaining information about the author to be used for marketing and product usage, building adaptive intelligent systems to better suit the product to their users, and determining authors of unknown texts for authenticity or other reasons.

Context vectors help in representing word tokens based on their neighboring words and their usage, this helps in capturing useful information about the style of an author which is often lost when using other word embedding methods like Word2Vec.

Statistics from a similarity matrix generated after comparing the context vectors of tokens are used to capture stylistic features. For e.g, the style of an author who uses a particular set of vocabulary or unique words in relation to each other can be captured by these statistics.

This paper explores the amount by which the style of an author determined by context vectors can influence the task of authorship verification. Usage of context vectors has not been done before in this task. The paper proposes to use context vectors alongside other commonly used features like word n-grams, character n-grams, and others using various kinds of similarity measures and configurations.

It was found that using context vectors helps in increasing the similarity predictions of document pairs. The top n Frequent tokens perform the best followed by singleton tokens and using every token. Cosine similarity has also been found to perform better than euclidean distance and The imposter strategy also gives lower absolute error than the universum strategy.