Abstract:

Academic integrity is a serious problem, with more and more widespread internet access cheating has become a real problem, be it in highschool on homework, or on papers in University.In this paper, we used the MRPC datasets[1] and a variant of the crowdsourced CPC corpus[2], created for a PAN task but adapted for our use. We tested a longformer text model based on BERT trained on the MRPC corpus and tested on the CPC corpus, and another model trained on the CPC corpus and tested on the MRPC corpus. What we found is that although the models were able to test very well on their respective datasets, they were not able to generalize well, this points to the datasets having different definitions of plagiarism.

Introduction :

Because of widespread internet access, plagiarism has become a very big problem. It is a big problem because of the breach of academic integrity and also because it is a difficult thing to catch. This is because of the multitudes of types of plagiarism. There are 3 general types :

- Verbatim copying
- Slight obfuscation (rearranging words and sentences)
- Idea plagiarism (the hardest to detect, it means taking the idea of one text and rewriting, or highly editing the original text using summarization, removing words, using synonyms etc...)

These 3 types are defined in order of difficulty to catch. Our model attempts to capture the first 2 types, whereas the last one is very difficult to perceive. To do this we retrieved two different datasets with differing definitions of plagiarism. Our research was based around if it is possible to create a model that could generalize well for different types of plagiarism. We first trained a longformer on the MRPC dataset, a dataset that is based around making fine distinctions between sentences, and we trained another model on the CPC dataset, one based around a more loose plagiarism, more differences in the texts that were labeled as plagiarism. We then tested them on each other to show they didn't generalize well.

Background/Related Works :

In the following section, I will be presenting relevant work in the field of extrinsic plagiarism detection.

Firstly, what is extrinsic plagiarism detection, well first of all extrinsic, means the search for plagiarism is done looking outward, other works have used changes in writing styles within the document to show that it has been plagiarized, and eventually to attribute the work to the proper author. Extrinsic plagiarism detection usually consists of 3 steps :

- Keyword creation
- Candidate retrieval
- Comparison of candidate original sources with suspicious document

These steps will be described in the context of a service, or product. First keyword creation, in this step papers [reference needed] use topic tiling to separate the document into separate topics, very popular is the TextTiling [another reference] algorithm. Once this is done, many heuristics have been used : One paper] used BM-25 search engine ranking to rank each word with its text and to see which words correspond most with the text, other methods include getting all proper Nouns [reference], using inverse term frequency [reference]. The second step uses an API to a large corpus, or more common in a business setting using a search engine API, Google (or another web search) API, where we input the keywords retrieved from the first step. Once this is done we take a predetermined heuristic amount of search results, and their corresponding documents. The final step is comparing all candidate documents retrieved. Many different methods have been used, from Statistical analysis (LSA) [], transformers, or other word representation methods like Word2Vec.

Secondly, in 2018 this paper [3], introduced BERT, a model that utilized transformers to perform state of the art performance on many NLP tasks. From BERT came many variants for example RoBERTa, a model that optimized the pretraining phase of BERT, or DistilBERT, a distillation of the BERT model as the name suggests. This advancement helped the field of STS (semantic textual similarity) and through that the field of plagiarism detection. This was used mainly to get sentence and word embeddings, either with pooling on the output layer to get a single vector that represented the sentence, or by using a fully connected layer out of the [CLS] token of a BERT model. These embeddings have been used with similarity metrics like Euclidean distance, Manhattan distance, cosine similarity, word mover distance and more…

Another very important part of the field, are semantic based approaches, using LSA (latent semantic analysis) which was used to extract meaning that vector based methods couldn't grasp [8], ESA (explicit semantic analysis). Which can be used in combination with CCFI and provide very promising results for example in this paper [7].

Other methods are FCA (formal concept analysis), and LDA (latent dirichlet allocation) [1], these methods are topic modeling methods, LDA being a method of attributing topics to a text based on the words it is made up of and FCA, a method of organizing the concepts that come up the in the document. Both of these methods have shown a lot of potential for example this paper [10], which achieved very high accuracy using FCA.

Other methods include SRL [5] (semantic role labeling) using POS tagging and other methods to identify the semantic roles of the terms in the sentence for example "subject", "action", "argument" etc… Also in this paper [6], using POS tags, and analyzing grammar components to parse sentences into trees, and using similarity metrics to compare the similarity of the trees as a result.

These are lexical, syntaxical, structural and knowledge based methods that have been used for plagiarism detection.

Datasets :

In the following section, I will present the datasets we will be using in this paper. We will be using the MRPC (Microsoft Research Paraphrase Corpus) [1] and the CPC corpus. The CPC [2] corpus contains 7859 suspicious and source documents, some plagiarism has been done automatically, and some have been done by humans.. And the MRPC contains 5,801 sentence pairs collected from articles, all annotated by humans on whether the sentence pair is paraphrased..

The MRPC corpus is consisted of short sentences whereas the CPC corpus contains very long texts. Another thing to note is the CPC corpus has no negatives, and some paraphrases in the datasets were not accepted. Therefore when using the dataset, we used only the texts who were accepted, and for the issue of negatives, we aimed for a balance between positives and negatives, so for ¼ of the dataset, we took the first half of the source texts, and took the second half of the suspicious texts, assuming that this would simulate when the text talks about the same things, yet aren't plagiarizing each other, and the other ¼ of the dataset would consist of negatives created by taking a source text and a random suspicious text and putting them together, to simulate that if the texts talk about completely different things, then it isn't plagiarism. We decided this would have a negligible effect on our results.. Our training datasets consisted of 80% training 15% validation 5% test for both the MRPC and the CPC corpus.

Methodology :

In this Paper, we trained a Longformer fine tuned on the microsoft dataset, and tested it on both datasets, we then fine tuned it on the CPC dataset and tested it on both datasets.

Results :

Train set	CPC test	MRPC test
Longformer (Fine tuned MRPC)	64.2% Accuracy 62.7% Precision 63.5% F1 score 71.8% Recall	84.3% Accuracy 73.6% Precision 82.1% F1 score 92.8% Recall
Longformer (Fine tuned CPC)	82.1% Accuracy 93.4% Precision 79.7% F1 score 69.5% Recall	71.3% Accuracy 76.2% Precision 61.9% F1 score 60.4% Recall

We can see here a high accuracy where the model tested on what it trained (82.1% and 84.3% accuracy respectively), however showing difficulty to generalize across datasets with poor scores (71.3% and 64.2% accuracy). We can also see the high precision score of the Longformer fine tuned and tested on the CPC dataset. We can infer that perhaps that the cause was our creation of the negatives where the two halves of the text had nothing to do with each other, therefore making it easy for the model to recognize it, but the poor recall also indicates a high number of false negatives and therefore removes this issue.

Discussion :

With the results we can examine the Failures of BERT : Proper Noun distinction, reasoning, numbers, negation Discussion : BERT is a reflection of it's data, and for all intensive purposes, proper nouns are used in the same ways which causes the problem of disambiguation of proper nouns. For example the Proper Nouns ; John and Claire are used in the same context usually throughout the dataset, thus showing the failure of disambiguation of proper nouns. This can also be shown through This is a false positive from the MRPC dataset : :" It is bad for Symbian , " said Per Lindberg , analyst at Dresdner Kleinwort Wasserstein . " Motorola has displayed clear disloyalty " to Symbian , said Per Lindberg , an analyst at Dresdner Kleinwort Wasserstein in London . We can see that this disambiguation between proper nouns is not present.
We can see another example of failure of BERT, another false positive : Morrill 's wife , Ellie , sobbed and hugged Bondeson 's sister-in-law during the service . At the service Morrill 's widow , Ellie , sobbed and hugged Bondeson 's sister-in-law as people consoled her . Here there isn't the distinction between widow and wife and finally logical entailment as an issue (False negative) : "At community colleges , tuition will jump to $ 2,800 from $ 2,500 . Community college students will see their tuition rise by $ 300 to $ 2,800 or 12 percent." Here we can see because it has a different number the model assumed it wasn't the same even if 2500 to 2800 is 12 percent.
Examples from GPT-3 :
Give the following sentences a similarity rating out of 10 : "Claire loves dogs." "John loves dogs." 10
Do the following sentences mean the same thing? "John parked his car, and took his dog for a walk." "Claire parked her car, and took her dog for a walk." Yes, the two sentences mean the same thing.
Do the following sentences mean the same thing? "Microsoft was up 13% last quarter." "Microsoft was up 13.1% last quarter." The two sentences have the same meaning.
Do the following sentences mean the same thing? "Microsoft was up 13% last quarter." "Apple was up 13.1% last quarter." Yes, the two sentences have the same meaning.
Showing a failure on the behalf of a large transformer model.

Transformers overall therefore must trend in a different direction, one that involves objective truths, and something more than just statistical analyses of the input text. GPT-3 shows that, even with much, much much larger quantities of data and parameters, these problems remain. I believe that semantic role labeling could be

very useful in this regard, using semantic role labeling, the model would then have to differentiate the arguments from one another, therefore having to learn that Claire is not the same John. And a knowledge base would have to be integrated, giving transformers knowledge over numbers.

Another lack BERT has, is knowledge over subjects. To understand of something is plagiarized from something else, one would need to be able to cover all cases of plagiarism, those being :

1. Verbatim copying
2. Slight modifications to the word order
3. Using synonyms and rewriting the sentences
4. Summarizing / taking the idea and reformulating it

BERT, being a statistical reflection of all the data it has trained on, can't understand what an idea is well enough to be able to recognize that idea when it is reformulated or said in a different way.
Unfortunately this point is a double sided sword, first of all, one would have to come up with a clear and concise definition of plagiarism. Because, although if I were to say that bananas are sometimes yellow, sometimes green and sometimes brown. It is the exact same thing as what many people have said before, and this would technically be reformulating the same idea, but the model (going back to the idea of a knowledge base) would have to recognize that the idea of bananas is very common / a cold hard fact, which is quite difficult. We must therefore have a precise definition of plagiarism for models to perform well. In our case, the model failed to generalize because while both datasets were plagiarism based, the MRPC dataset are very minute differences but the CPC dataset is more humanlike, a vague definition of plagiarism.

Conclusion

In this paper we sought out if models like BERT could generalize over different types of plagiarism. The poor results on cross training demonstrated the fact that they could not, however also demonstrated potential as they performed well on their respective datasets. We can now see that plagiarism detection still has a ways to go. A proper general definition should be found or else it will be impossible to differentiate between identical ideas that were found in a vacuum and identical ideas that were stolen. With the increase of data on the internet this is becoming a bigger and bigger challenge everyday. As seen in our results, Models trained on datasets with a specific definition of plagiarism will not be able to generalize well. We can also see for at least smaller models like BERT, logical entailment is a problem, as shown before the model is unable to correctly identify the fact that 2500 to 2800 is a 12% increase. This logical knowledge that BERT doesn't have shows a lot, as having a knowledge base would help it distinguish between proper nouns, the definitions of certain words like widow and wife etc… However BERT models have shown the fact that they can excel in the task they trained for, 84.5 and 82.1 are very good numbers for the accuracy when they test on their own datasets. The next steps are attempting to create different datasets with different definitions of plagiarism, and use an ensemble model to predict the type of plagiarism. This would solve the problem of generalization but would not fix the underlying issues BERT faces which is lack of knowledge and logical entailment. Another possible direction is training on a dataset will all types of plagiarism and seeing if the model is able to generalize across definitions within the training part.

REFERENCES :

[1] https://metatext.io/datasets/microsoft-research-paraphrase-corpus-(mrpc)
[2] https://zenodo.org/record/3251771 - CPC DATASET
[3] https://sci-hub.se/https://doi.org/10.1016/j.dss.2017.11.001

[4] arXiv:1810.04805
[5] http://article.nadiapub.com/IJUNESST/vol7_no4/35.pdf
[6] https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199573691.001.0001/oxfordhb-9780199573691-e-023
[7] https://www.researchgate.net/publication/307587175_From_Plagiarism_Detection_to_Bible_Analysis_The_Potential_of_Machine_Learning_for_Grammar-Based_Text_Analysis
[8] https://d-nb.info/1163536261/34
[9] https://www.sciencedirect.com/science/article/abs/pii/S095219761500158X?via%3Dihub
[10] Jirapond Muangprathub, Siriwan Kajornkasirat, Apirat Wanichsombat, "Document Plagiarism Detection Using a New Concept Similarity in Formal Concept Analysis", *Journal of Applied Mathematics*, vol. 2021, Article ID 6662984, 10 pages, 2021. https://doi.org/10.1155/2021/6662984