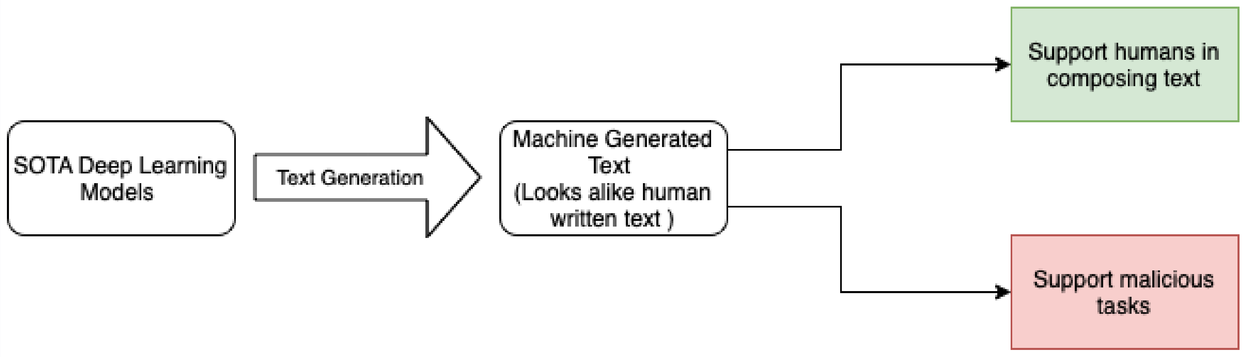
A Benchmark Corpus for the Detection of Automatically Generated Text in Academic Publications

Vijini Liyanage, Davide Buscaldi, Adeline Nazarenko

LIPN, Université Sorbonne Paris Nord

Is automatic text generation always used in a responsible way?



Examples in academia

- Machine-assisted plagiarism
- Artificially increasing the number of publications
- Citation index boosting (creating publications that cite one's own works)
- Fake reviews

Detection of automatically generated texts, a difficult task

Human detection

- "The accuracy of human detection of artificially generated text without any tool is only 54%." [1]
- It improves only slightly when assisted with GLTR [1]

Automatic detection

- GROVER [4] can only predict those texts generated by the model itself
- RoBERTa [2] has not been tested on academic texts

A benchmark corpus consisting of two datasets

Fully generated dataset

- Automatically generated academic papers using GPT-2 model
- 100 articles (average length: 1243 words)

Hybrid dataset

- Abstracts partly written by humans and partly generated artificially using the Arxiv-NLP model
- 100 abstracts (average length: 177 words)

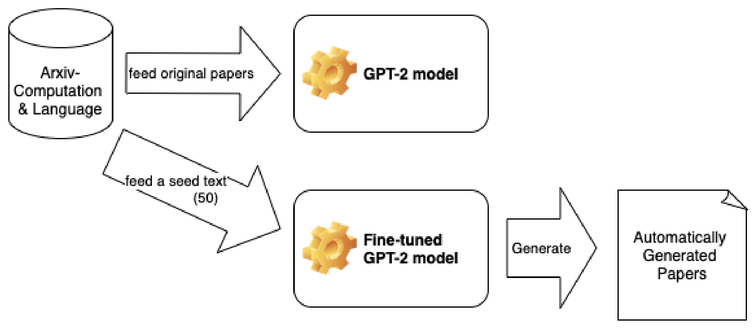
Original versus generated content examples

Fully generated dataset	
Original Paper	Generated Paper
Finally, in this task, we used comparative evaluations to **determine** the quality of the sets of generated **KPs. Obviously, this results with ranking based evaluation. Establishing an absolute evaluation metric in this context merits further investigation. We thank all participating teams for taking the time to participate in this challenging new shared task. We believe that these contributions, along with the data collected and shared in this report, will be valuable to further advance the research around KPA and related tasks.**	Finally, in this task, we used comparative evaluations to **estimate** the quality of the sets of generated **texts. We expect the task and the findings reported in this paper to be relevant for researchers working on text summarization and argument mining. Introduction Key Point Analysis (KPA) was introduced in BarHaim et al. (2020a,b) as a challenging NLP task with tight relations to Computational Argumentation, Opinion Analysis, and Summarization, and with many practical applications (Bar-Haim et al., 2021).**

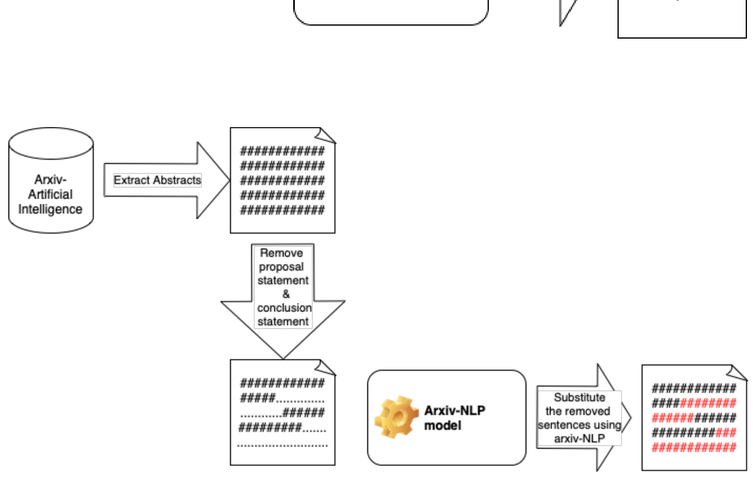
Hybrid dataset	
Original Abstract	Generated Abstract
Our experiments suggest **that models possess belief-like qualities to only a limited extent, but update methods can both fix incorrect model beliefs and greatly improve their consistency.**	Our experiments suggest **the importance of model beliefs in learning models, and we show that the approach outperforms automatic model updating systems using word representations.** .

Corpus design methodology

Fully-generated Dataset



Hybrid Dataset



Benchmark corpus evaluation based on two criteria

1. How natural the generated content looks like?
   - BLEU score
   - ROUGE score
2. How difficult is it to distinguish between generated and human written content?
   - Classification accuracy of statistical models (Naive Bayes, Passive Aggressive, Multinomial with hyper-parameters & SVM)
   - Classification accuracy of deep learning models (LSTM, Bi-LSTM, BERT & DistilBERT)

BLEU & ROUGE Score results

Fully-generated Dataset					
Score	unigram level BLEU	sentence BLEU	Rouge-1	Rouge -2	Rouge-L
Min. Score	0.659	0.583	0.591	0.559	0.591
Avg. Score	0.867	0.809	0.853	0.810	0.853

Hybrid dataset					
Score	ngram level BLEU	sentence BLEU	Rouge-1	Rouge -2	Rouge-L
Min. Score	0.629	0.495	0.598	0.473	0.598
Avg. Score	0.824	0.792	0.882	0.840	0.881

- The artificial texts are quite similar to natural ones (average scores > 0.8)
- This attests the robustness of the dataset

Classification results

Fully Generated Dataset	
Model	Accuracy
Bag of words, Multinomial Naive Bayes Algorithm	19.7
Bag of words, Passive Aggressive Classifier Algorithm	31.8
Bag of words, Multinomial Classifier with Hyperparameter (alpha)	19.7
Bag of words, SVM	37.9
LSTM model	59.1
Bi-LSTM (Latest Paper)	40.9
BERT	52.5
DistilBERT	**62.5**

Hybrid dataset	
Model	Accuracy
Bag of words, Multinomial Naive Bayes Algorithm	24.2
Bag of words, Passive Aggressive Classifier Algorithm	30.3
Bag of words, Multinomial Classifier with Hyperparameter (alpha)	22.7
Bag of words, SVM	37.9
LSTM model	50.0
Bi-LSTM (Latest Paper)	47.0
BERT	50.0
DistilBERT	**70.2**

- The deep learning based models achieve higher accuracy scores than the statistical ones
- The highest classification score is obtained by the DistilBERT model
- The accuracy scores are not very high, even for DistilBERT
- This shows the quality of our benchmark in terms of classification difficulty

Comparison of classification results wrt. a SOTA corpus

Model	Fully generated dataset	Hybrid dataset	Dataset of Maronikolakis *et al.*, 2020
Bi-LSTM	40.9	47.0	82.8
BERT	52.5	50.0	**85.7**
DistilBERT	**62.5**	**70.2**	85.5

- Our classification accuracy scores are lower than those obtained by [3] on a different dataset[1]
- This shows that our datasets are more difficult to classify
- This makes our corpus a better benchmark proposal than previous ones

Conclusion

- We have fine-tuned text generation SOTA models to produce coherent, similar to human-written academic content.
- We have built two datasets to experiment detection of automatically generated text.
- The evaluation shows the quality of the corpus both in terms of the naturalness of the texts that make it up and in terms of the difficulty of distinguishing those that are machine-generated.
- The results show that the existing state of the art models for classification provide a maximum accuracy of 70.2% on our dataset.

Future Work

- We aim to increase the size of the datasets by adding more papers to the corpus.
- We plan to analyse the classification errors to identify some strategies to improve existing classification methods.
- We hope to leverage other SOTA generation models and/or incrementally improve them to enrich the dataset.

Bibliography

[1] S. Gehrmann, H. Strobelt, and A. Rush.
Gltr: Statistical detection and visualization of generated text.
arXiv preprint arXiv:1906.04043, 2019.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov.
Roberta: A robustly optimized bert pretraining approach.
arXiv preprint arXiv:1907.11692, 2019.

[3] A. Maronikolakis, H. Schutze, and M. Stevenson.
Identifying automatically generated headlines using transformers.
arXiv preprint arXiv:2009.13375, 2020.

[4] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi.
Defending against neural fake news.
arXiv preprint arXiv:1905.12616, 2019.

1 This may be due to the fact that they [3] focuses on the generation of short content (headlines)