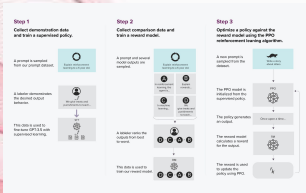
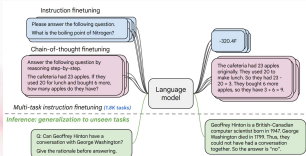


Fine-tuning Transformer-based Natural Language Generation Algorithms for USDA Grains Reports for Farmers, Producers, and Small Businesses

Winston Zeng

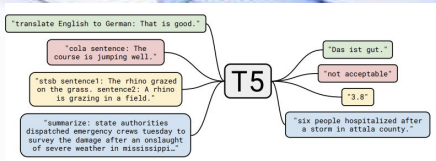
Introduction

The Transformer architecture for Natural Language Generation (NLG) was nothing short of a revolution for Natural Language Processing. Self- and multi-head attention models have proven their efficacy in a variety of textual tasks, including classification, translation, summarization, and generation. Fine-tuning Transformer-based models on summarization tasks has proven successful with small textual datasets. In this research, we focus on fine-tuning pre-trained transformer-based NLG algorithms for USDA grain reports to produce high-quality summaries, highlights, narratives, and Q&As, to enable farmers, producers, and small businesses in making informed decisions on production, investments, expansions, and risk management. For my thesis, I conducted a comparative analysis of different high-performing AI models, pre-trained on large textual datasets such as news article datasets, others publicized and popularized for their ubiquitous good performances.



Example Summary Pair (Good and Bad)

The US 2022/23 cotton supply and demand forecasts are unchanged, with a projected price of 83 cents per pound. The global outlook shows higher production and stocks, with beginning stocks nearly 900,000 bales higher due to updated consumption estimates for China and Uzbekistan. World cotton consumption is 555,000 bales lower with reductions in Turkey, Pakistan, Indonesia, and Bangladesh, and lower projected imports for these countries, and for China. Exports are also lower for Brazil, India, and Argentina, with world trade totaling 785,000 bales lower. Production is over 700,000 bales higher due to larger crops in China, Australia, and Uzbekistan, despite reduced prospects for India. 2022/23 world ending stocks are projected 2.1 million higher than a month earlier at 91.1 million bales, and 5.0 million higher than in 2021/22.



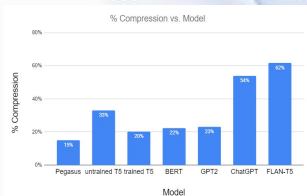
Model: ChatGPT

Compression: 75%

This month's 2022/23 U.S. cotton supply and demand forecasts are unchanged relative to last month. The global 2022/23 cotton supply and demand forecasts this month include lower consumption and trade, and higher production and stocks.

Model: BERT

Compression: 20.6%



Conclusion

- ChatGPT has the best overall performance, but experiences mild wordiness
- FLAN-T5 was the second best, but tends to omit information that comes later in the original report, as well as having to manually tinker with the output length constraints
- Untrained T5 was a surprising third overall best, as well as being the easiest to implement and maintain
- Training a T5 model hurt its performance when it was trained on news reports that were made up of mostly words as opposed to numerical information
- Overall, numbers tend to throw off abstractive summarization quality unless the model knows to specifically include the numbers while paraphrasing everything else

