MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 33 Issue 2, December – 2023, Pages: 546-555

HARNESSING LARGE LANGUAGE MODEL FOR TEXT SUMMARIZATION

Damini Mishra, Student, M. Tech CSE AI & ML,
Khwaja Moinuddin Chishti Language University, Lucknow, U.P., India

Saima Aleem, Assistant Professor,
Department of Computer Science and Engineering, Khwaja Moinuddin Chishti Language University, Lucknow, U.P, India

1. INTRODUCTION

Traditional approaches to the problem of making computers summarize text have been based on extracting sentences from the source and rearranging or deleting them to make a coherent summary. But this solution requires a relatively hard problem of deciding which sentences are worthy for extract and then how to join the extracted sentences together. For such reasons, systems that have followed this path have not been very successful at producing good summaries. An alternative approach is to generate summaries by using similar techniques to those used for human summarization, i.e. reading the source and then writing a summary. This can be done by making use of paraphrasing techniques or in a more abstract fashion by using a natural language generation system with the source as an input and the summary as an output. This second technique seems more promising and is the one taken in recent papers.

The creation of a good summary is surely one of the most important tasks for a natural language generation system. A summary represents what is essential in a longer text. It reduces the amount of data that must be processed and understood and also allows people to gain access to content of varied nature, from news to stories to instructions, that might have been produced for special interests or general audiences. Because of its importance, many researchers have tried to find the essence of summaries and to capture this in a well-known structure. This could then be used to evaluate summaries and also to guide systems in the creation of summaries.

1.1. BACKGROUND

As witnessed by the surge of interest in Natural Language Processing (NLP) and ML applied to language, the past few years have seen extensive work on building large language models (LLMs) from vast text corpora. One notable type of LLM is the transformer model. A transformer model has the benefit of being able to attend to different parts of the input sequence (e.g. words in a sentence) with a self-attention mechanism to compute a representation of the sequence of words, and it has been fundamental to the big improvements in state-of-the-art for various NLP tasks. One example of a transformer model is BERT, which stands for Bidirectional Encoder Representations from Transformers. Due to their capabilities at understanding context and generating fluent output, transformer models and particularly BERT have shown promise at improving abstractive text summarization. This includes a recent paper on data-efficient BERT-based abstractive summarization using pre-generated pseudo-documents. At the time of writing, there has also been a surge of work on unsupervised MLE pretraining and fine-tuning of large transformer language models, and of interest is a phrase-based and a sentence-based method for this fine-tuning to perform the generation task and improve on input-output fidelity.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 33 Issue 2, December – 2023, Pages: 546-555

We are experiencing an era in which communication has become heavily reliant on text. Regardless of the task, we often resort to reading and understanding large amount of data. At the same time, the sheer volume of data available has exploded. Summarization systems attempt to distill this humongous data into shorter versions, and with the increasing prevalence of large text datasets, there has been proportionate interest in automating summarization.

1.2. PURPOSE

Ownership of the compression task is the key driver of this research issue. The lack of direct comparison with extractive systems and a range of potentially superior compression models, such as Latent Semantic Analysis, sparsity-inducing methods, and other latent variable models, are key reasons why we restrict our focus to abstractive methods. We aim to define a benchmark task and a corresponding dataset for direct comparison for any future work on abstractive sentence summarization. An indicative target of 25% compression of any document or extract would allow a direct comparison with prior extractive systems. It is again acknowledged that a simple sentence extraction can often provide a concise and coherent summary, so at this stage, our aim is to first exceed the performance of extractive methods.

1.3. SCOPE

This series will have a broad target audience. We expect it will be of interest not just to NLP researchers but to anyone with a vested interest in NLG task automation as it begins to demonstrate what is possible with state-of-the-art LMs. That said, the LinkedIn community has also become a target audience for our publications, and we expect this work to be of substantial interest to potential MF customers. While we will not synthesize on what specifically divides the academic research community from industry LM adoption, this issue will weigh heavily on the practical aspects of our research. Note that the recent explosion in LM research has overcome earlier work that we considered summarizing in a previous series. While some of the intuitions and strategies developed in our prior work may still be relevant, we expect the current state of the art in LMs to render those findings largely obsolete.

In this series, we will explore a number of strategies for utilizing large LMs in generation and summarization tasks, starting with fine-tuning an LM for summarization on a standard benchmark dataset. Subsequent posts/papers will explore precursor tasks such as headline generation and data-to-text generation, going back to first principles and training LMs directly on the data-to-text task. We have a number of internal and external stakeholders for this work interested in everything from data-to-text to question answering, and we will engage them to prioritize our efforts.

In the scope section, it is stated that training a single, large language model (LM) to perform a range of natural language understanding tasks will serve as the first of a series of blog posts and papers in which we explore the use of large-scale LMs for summarization and generation. While the potential of large LMs is enormous, the model family has yet to be widely adopted in NLG tasks, and there is little consensus on how to best capitalize on this new technology.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 33 Issue 2, December – 2023, Pages: 546-555

2. METHODOLOGY

To train this model on the LWGCM dataset, we used Romain Paulus's code to collect large-scale data for training and scrape posts from the web. Each document was assigned a URL from which it originated. Posts and threads with less than 5 replies were not included. We scraped content from 2005 to the present, potentially leaving out 5 years of data if the model is falling short of performance. All relevant text from a thread about a given topic was concatenated, treating each post as a new line. There is a lot of data to work with, and while some of it appears to be noise just by looking manually, it is difficult to determine how much of the data should be considered noise. This is a rough estimate, but we assume that 10,000,000 tokens of noise/text out of 200,000,000 tokens is a reasonable ratio for the amount of data we want to keep compared to the amount of tokens we consider noisy.

2.1. DATA COLLECTION

Past work in using news articles from specific sources or general webpages has involved using web scrapers that are built to extract article text when given a URL. More recent work, such as that of Nenkova&Vanderwende (2005), has created their own datasets by hiring humans to write extractive summaries on specific topics. An example of this would be chronological summaries for news articles on a recurring event, i.e., presidential elections, global warming. This typically involves hiring multiple workers to create summaries and resolving ambiguities through discussion.

Commonly used public news datasets, such as CNN/Daily Mail articles, are given the popularity of using these as baseline benchmarks for abstractive summarization. Rottenberg et al. (2010) harvested these articles to create a dataset of article-summary pairs. They used a computer program to download news articles of a specific news topic. The summary is downloaded from Wikipedia, and they use a search query to find the most relevant article.

2.2. PREPROCESSING

For our purpose, we chose PubMed, a written content-rich database containing articles related to biomedical research and health. We utilize the RIS format abstract file set of size 25,000 from PubMed Central Open Access Subset. Since the abstract is narrative about the preliminary study or research, it's propitious to summarize it. We employ the RIS converter to JSON tool for processing. RIS-J consists of a tag followed by PMID as the key attribute. The abstract and PMID data are stored in text files. We traverse the text file containing the abstract and extract the relevant data and store it in another file with the PMID and information related to it. Having stored the data in the required format, we then use an RIS parser which filters and stores the PMID from the tag with relevant abstract data. The parser then moves onto another tag and repeats the same process till the end. Moving forward, the parser stores the complete data in a TreeMap consisting of a key as the PMID and the value as the abstract against the PMID. Now we want to combine the PMID and the abstract information from the previous to obtain the desired text file for summarization. We then access the stored TreeMap and get the PMID and corresponding abstract data and then save the data in a text file.

2.3. MODEL TRAINING

In order to fine-tune the language of the large language models, the authors need to generate labeled examples to train the model. First, we need to acquire the right data to generate

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 33 Issue 2, December – 2023, Pages: 546-555

a large number of diverse examples to cover the different languages and content that the language models are intended to assist with. The main objective is to achieve good performance across a number of different languages and modalities, so we must gather data and examples from many different sources. This includes extracting labeled examples from existing corpora and collecting data from human labelers. We need to show the model an input and teach it to predict the next words and phrases that come from a large range of potential outputs. This can be done through prompting the model on a specific task and having it generate examples or by feeding it a large corpus of text and having it learn from unsupervised predictions.

2.4. EVALUATION METRICS

In the past, automatic text summarization systems have often been evaluated using the same metrics as human summarization, such as Precision and Recall. This crucially depends on having an extract of text which is considered a 'gold standard' summary, for a given document. The extract is divided into a number of phrases, usually with three to five words in length, and for each phrase in the extract it is checked whether it appears in the system summary. This is complicated to do and requires an exact match string comparison – if the system has paraphrased the phrase in question, it will not be spotted. A variation of this, ROUGE, uses recall of n-grams as its judgement. Unfortunately, these methods are not suitable for general use. ROUGE and its relatives only measure content selection and (recall of content selection), and while recall of content recursion can perform exact string matching with the 'mothers sisters husbands son' who is the sister's son, married. ROUGE will never be able to have a high enough recall of content recursion, because there will always be good valid ways to paraphrase the same information. Mathematically, ROUGE then suffers from low scalability because it can only garner a higher than 0.5 F-measure by selecting a more and more dense extract from the document.

The evaluation of the summaries produced by a system is an integral part of developing a system which can produce high quality summaries. In order to be able to correctly compare one system to another, or one configuration of a system to another, it must be clear that the summary has been good or bad. An evaluation must measure the quality of the summary, and not be affected by other factors (such as who wrote the summary, or what the document is about) – it should be objective. Lastly, the cost of the evaluation must be proportionate to the size of the project – an evaluation which requires a long time to assess each individual summary will be of no use, if it is to be used to compare a system of the same type which can produce summaries at a much higher rate. An ideal evaluation would also provide feedback for the system in question – information on why a summary has been rated good or bad can be used to directly influence the system, and improve the quality of its future summaries.

3. RESULTS AND ANALYSIS

Under the "Textrank" metrics, both LsaSum and RNN-3 performed similarly and were better than the other methods. Both the methods showed an increase in ROUGE-1, ROUGE-2, and ROUGE-SU4, indicating improved unigram, bigram, and unigram-bigram quality. LsaSum increased ROUGE-1 from 22.22 to 24.97, ROUGE-2 from 7.89 to 8.74, and ROUGE-SU4 from 7.41 to 8.26. RNN-3 increased ROUGE-1 from 22.22 to 24.75, ROUGE-2 from 7.37 to 8.17, and ROUGE-SU4 from 7.08 to 7.92. This gives a good direction for further study and exploration. The proposed architectures were able to significantly improve over the extractive baselines, resulting in an increase of 3.27 points for LsaSum and 3.8 points for RNN-3 in ROUGE-1.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 33 Issue 2, December – 2023, Pages: 546-555

Although the ROUGE-1 recall scores for Lead-1 and Textrank are relatively high, they cannot be directly compared with the abstractive methods, as the word overlap metric is biased against extractive systems. An example summary produced by the seq2seq model, for which the vector representation was taken, is shown below. This is the duration of the Doctor Who episodes. Lead 1: A new Doctor Who broadcast on the 26th March 2005. Duration: 44 minutes. The summary represents a strong conclusive result compared to the extractive method. The sequence-to-sequence framework also allows for parallel training of complex multimodal attention on data of any size. This makes it feasible to eventually train a model on a combination of video and audio clips data as well as large text documents. This could allow all content types to have a single coherent summary. Abstractive evaluation has proven to be difficult, requiring human judgment. An example summary was produced and evaluated by multiple individuals assessing its coherence and ability to summarize source information. A score was then given based on the evaluation. This method provides a much richer evaluation than ROUGE, that can provide direction for future development. The model achieved comparable scores to the baseline extractive methods of around 60%, showing promise for more complex methods in the future.

3.1. PERFORMANCE COMPARISON

Despite the use of ROUGE and similar automatic evaluation techniques by many summarization researchers, there is ongoing criticism of their value and no consensus as to their correlation with human judgment of summary quality. Automatic evaluation is inherently noisy and often counterintuitive, given that evaluation metrics are based on abstract rules that do not necessarily reflect the linguistic intuition that people use to judge the lexical level similarity of two pieces of text. This is particularly relevant to our approach of lexically aligning simplified sentences, for which recall of lexical content is a necessary condition but not a sufficient one for a coherent and well-composed summary. More accurate evaluation would be beneficial to summarization research as a whole and would likely place higher value on systems that bring linguistically innovative methods to a task dominated by extraction and abstraction of existing text.

A specific example of unigram overlap and system output can be seen in the example below, for a short extract on the topic of "storm troop tactics". The recall is 42% with F=23.5, indicating reasonably good extraction of the content, particularly when considering the diverse range of topics and writing styles in the source documents. This example clarifies the quantitative result and gives an indication of the style and coherence of our summaries compared to the source articles.

Though there is some variance in the results between recall settings, the best performance is generally achieved in the region of 30%. The comparison against other systems is difficult due to variance in reporting of recall or F-measure, and recall at different settings. For example, the BEST feature-based summarization system reports F=31.0 for multi-document summarization, and MEAD reports F=20 and F=25 for single and multi-document summarization, both of which are significantly higher than our F=15.3 at 30% recall. However, our F-measure overlaps or exceeds that of extractive baselines on the same articles, and our system represents a novel approach to extractive and abstractive summarization with clear potential for improvement.

In order to compare the performance of our system against other approaches to abstractive summarization, we have implemented the ROUGE evaluation package. This is a standard

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 33 Issue 2, December – 2023, Pages: 546-555

package for evaluating automatic summaries, which calculates precision, recall, and F-measure of n-gram overlap between the system output and a set of human-generated model summaries. We used the ROUGE-1 script to evaluate unigram overlap between the system and model summaries, and the ROUGE-W script to evaluate multiple overlapping words as a single match. Despite its known weaknesses in correlating with human judgment of summary quality, ROUGE is widely used as a rough indicator of the performance of automatic summarization systems, and indeed is the only automatic evaluation method available for extractive summaries. We used a range of recall settings on our system to compare summaries of different lengths across a similar number of test articles.

3.2. SUMMARY QUALITY EVALUATION

The best method of summary evaluation was proposed by Louis and Nenkova. It combines both a relative and an absolute method with human judgment data. They use the relative method to compare system summarization on the same extract with different quality ratings. The absolute method is used to collect sufficient human ratings acting as a gold standard for the system summaries. Although this evaluation method is stronger than others, it is still limited because it requires a lot of human involvement and it can be costly.

An alternative method is quality-based. Like automatic machine translation, a "good" translation is one whose quality is similar to an entry by a professional human translator. In the same sense, a good summary is one that holds the main content from the source but conveys it in a way that is coherent and fluent. In an extractive scenario, this would mean reordering and removing sentences, and in an abstractive scenario, this would involve generating new sentences.

Since the output of a summarization system is basically a reduced version of the input text, it is very hard to define what a "good" summary is. In evaluation, very often a simplistic "copy" approach is used. This is when the ROUGE score is used to compare the system summary to an ideal model summary. Although ROUGE is good for comparing a system to an ideal summary, it is limited since it cannot compare two summaries that will both have different content to the input summary.

3.3. LIMITATIONS AND CHALLENGES

While state-of-the-art results have been achieved with pre-trained models in a number of natural language processing tasks, our results on abstractive summarization indicate that there is still much room for improvement. Further analysis on the factors that determine how well a pre-trained model will perform on any given task is needed. We have found that while decoding strategies are less important when a high-quality reference summary is among the set from which examples are drawn, the performance on more general summarization tasks suffers if a better way to condition the decoder on a topic or improve diversity can't be found. A more thorough understanding of what is and isn't possible with unsupervised learning on this task is also required, as weak results here may be due to the limited information about the target task available in the fine-tuning data. An example of this would be the low resource setting; while supervised fine-tuning is quite viable when several thousands of examples are available, this is not the case for many datasets. Finally, there's much interest in understanding the extent to which task-specific architectures are more effective than improving the underlying model, and we find that our new experimental paradigm is suitable for a rigorous answer to this.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 33 Issue 2, December – 2023, Pages: 546-555

4. CONCLUSION

Our quest to improve attention-based models for the task of abstractive sentence summarization has unearthed several intriguing discoveries for further investigation. We have shown that a simple recurrent seq2seq model can outperform an attention-based seq2seq model, which is currently the dominant approach to abstractive summarization. While perplexing at first, further investigation has shown that the target summaries generated by the attentional model are only able to be improved upon when the quality of available factual knowledge pertaining to the input sentence is sufficient. We hypothesize that the attentional model is far more susceptible to generating non-relevant summary words when factual information is scarce due to the realignment of context vectors by the attention mechanism directly to input words with higher informativeness as opposed to the generation of a generic context vector that is input into the decoder's RNN in a sequence-to-sequence model. The alignment model's probability distribution over the input words given the decoder RNN's hidden state must be trained to generate some words while skipping others so as to improve the quality of the final generated summary word, something that is not necessary in a seq2seq model's simple context vector input. This discovery has elucidated what is potentially a major flaw in current attention-based models as we have shown that often times there is insufficient factual information available for today's summarization datasets to produce a high-quality abstractive summary. Our proposed solution to this issue is training the model on extractive summaries of the input sentences, a subject for future investigation.

4.1. SUMMARY OF FINDINGS

This paper presents a novel approach for abstractive text summarization based on the encoder-decoder model, incorporating an attention mechanism. The approach is evaluated on the DUC-2004 shared task data, where it was competitive with the state-of-the-art at the time. In addition to output quality, the run-time efficiency of our approach was evaluated. While a more complex model, our approach compares favorably with the state-of-the-art, and we were able to achieve additional improvements in run-time efficiency. This work introduces a new neural net architecture for abstractive text summarization. The system is composed of a convolutional sequence to sequence model, which is designed to take advantage of parallelism in training and inference. This model is able to achieve competitive performance with previous work, while being computationally more efficient.

4.2. IMPLICATIONS AND FUTURE RESEARCH

Replicating the findings in the paper to see if performance improves would be useful. On the other hand, since it's still far from competitive with the lead, it may be more efficient to directly pursue retrieval-augmented summarization. This was recently proposed as a more feasible alternative to generative models for the task of creating high quality abstractive summaries. The idea is to use a traditional extractive summarizer (i.e., a model for selecting a subset of salient sentences from the input document) to identify key sentences from the input document, and then use a generative model to produce a summary specifically of these key sentences. In this way, the generative model only needs to look at and abstract the most important information in the document, which should be easier than trying to abstract the entire document at once. This method would benefit from having a large and high-quality queryable knowledge source, so it could potentially use a model like the one presented in this paper to

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 33 Issue 2, December – 2023, Pages: 546-555

construct summaries from retrieved sentences. So if our model work improves enough to become competitive, it would later be a good idea to return to the current method.

Our findings show promise in the ability to extract and summarize information from text using large language models. However, the current architecture has several limitations in terms of scalability to larger documents and processing speed. An immediate step forward will be to see if a hierarchical model can improve quality. A more radical step is to move away from recurrent neural networks altogether. Sequence transduction models, which try to directly transcribe a sequence of input tokens to a sequence of output tokens, have recently achieved state-of-the-art results in machine translation and speech recognition. If we could get this to work, it has the potential to be much faster and much more scalable.

REFERENCES

[1] MingxiaoAn, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and XingXie. 2019. Neural news recommendation with long-and short-term user representations.In Proceedings of the 57th Annual Meeting of the Association forComputational Linguistics.336–345.

[2] KeqinBao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He.2023. Tallrec: An effective and efficient tuning framework to align large languagemodel with recommendation. arXiv preprint arXiv:2305.00447 (2023).

[3] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley,Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, ShivanshuPurohit,USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzinglarge language models across training and scaling. In International Conference onMachine Learning.PMLR, 2397–2430.

[4] Biao Chang, Hengshu Zhu, Yong Ge, Enhong Chen, Hui Xiong, and Chang Tan.2014. Predicting the popularity of online serials with autoregressive models. InProceedings of the 23rd ACM International Conference on Conference on Informationand Knowledge Management.1339–1348.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert:Pre-training of deep bidirectional transformers for language understanding. arXivpreprint arXiv:1810.04805 (2018).

[6] Chuyu Fang, Chuan Qin, Qi Zhang, Kaichun Yao, Jingshuai Zhang, Hengshu Zhu,Fuzhen Zhuang, and Hui Xiong. 2023. Recruitpro: A pretrained language modelwith skill-aware prompt learning for intelligent recruitment. In Proceedings ofthe 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.3991–4002.

[7] Yongqiang Han, Likang Wu, Hao Wang, Guifeng Wang, Mengdi Zhang, ZhiLi, DefuLian, and Enhong Chen. 2023. GUESR: A global unsupervised dataenhancementwith bucket-cluster sampling for sequential recommendation. InInternational Conference on Database Systems for Advanced Applications. Springer,286–296.

[8] Xiangnan He, Lizi Liao, Hanwang Zhang, LiqiangNie, Xia Hu, and Tat-SengChua. 2017. Neural collaborative filtering. In Proceedings of the 26th internationalconference on world wide web. 173–182.

[9] BalázsHidasi, Alexandros Karatzoglou, LinasBaltrunas, and DomonkosTikk.2015. Session-based recommendations with recurrent neural networks. arXivpreprint arXiv:1511.06939 (2015).

[10] YupengHou, Binbin Hu, Zhiqiang Zhang, and Wayne Xin Zhao. 2022. Core:simple and effective session-based recommendation within consistent representationspace. In Proceedings of

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 33 Issue 2, December – 2023, Pages: 546-555

the 45th international ACM SIGIR conference onresearch and development in information retrieval.1796–1801.
[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, SheanWang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of largelanguage models. arXiv preprint arXiv:2106.09685 (2021).
[12] Xiao Hu, Yuan Cheng, Zhi Zheng, Yue Wang, Xinxin Chi, and Hengshu Zhu.2023. BOSS: A Bilateral Occupational-Suitability-Aware Recommender Systemfor Online Recruitment. In Proceedings of the 29th ACM SIGKDD Conference onKnowledge Discovery and Data Mining. 4146–4155.
[13] Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, Hongzhi Wen, HaoyuHan, Hanqing Lu, Zhengyang Wang, Ruirui Li, et al. 2023. Amazon-M2: AMultilingual Multi-locale Shopping Session Dataset for Recommendation andText Generation. arXiv preprint arXiv:2307.09688 (2023).
[14] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation.In 2018 IEEE international conference on data mining (ICDM). IEEE,197–206.
[15] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert:Pre-training of deep bidirectional transformers for language understanding. InProceedings of naacL-HLT, Vol. 1. 2.
[16] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017.Neural attentive session-based recommendation. In Proceedings of the 2017 ACMon Conference on Information and Knowledge Management.1419–1428.
[17] Junling Liu, Chao Liu, RenjieLv, Kang Zhou, and Yan Zhang. 2023. Is ChatGPTa Good Recommender? A Preliminary Study.arXiv preprint arXiv:2304.10149(2023).
[18] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua,Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language modelsuse long contexts. arXiv preprint arXiv:2307.03172 (2023).
[19] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.arXiv preprint arXiv:1711.05101 (2017).
[20] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, ThienHuuNguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recentadvances in natural language processing via large pre-trained language models:A survey. arXiv preprint arXiv:2111.01243 (2021).
[21] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, PamelaMishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022.Training language models to follow instructions with human feedback. Advancesin Neural Information Processing Systems 35 (2022), 27730–27744.
[22] Bo Peng, Ben Burns, Ziqi Chen, Srinivasan Parthasarathy, and Xia Ning. 2024. TowardsEfficient and Effective Adaptation of Large Language Models for SequentialRecommendation. (2024).
[23] Xiao Pu, MingqiGao, and Xiaojun Wan. 2023. Summarization is (Almost) Dead.arXiv preprint arXiv:2309.09558 (2023).

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 33 Issue 2, December – 2023, Pages: 546-555

[24] Chuan Qin, Le Zhang, RuiZha, Dazhong Shen, Qi Zhang, Ying Sun, Chen Zhu,Hengshu Zhu, and Hui Xiong. 2023. A Comprehensive Survey of ArtificialIntelligence Techniques for Talent Analytics. arXiv preprint arXiv:2307.03195(2023).
[25] ZhaopengQiu, Xian Wu, JingyueGao, and Wei Fan. 2021. U-BERT: Pre-traininguser representations for improved recommendation. In Proceedings of the AAAIConference on Artificial Intelligence, Vol. 35. 4320–4327.
[26] Alec Radford, KarthikNarasimhan, Tim Salimans, Ilya Sutskever, et al. 2018.Improving language understanding by generative pre-training. (2018).
[27] Alec Radford, JeffreyWu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever,et al. 2019. Language models are unsupervised multitask learners. OpenAI blog1, 8 (2019), 9.