JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

NLP LATE FUSION MULTI LINGUAL TEXT SUMARISATION

Dhanush S R
MSc DS *Student,Dept of CSA*
REVA University
Bangalore,Karnataka,India

DR.Manju Priya
Associate Professor,Dept of CSA *REVA University*
Bangalore,Karnataka,India
smanjupr@gmail.com

Abstract*—For a variety of cross-lingual transfer tasks, pretraining multilingual language models at scale results in notable performance improvements[1]. We use almost two terabytes of filtered CommonCrawl data to train a Transformer-based masked language model on one hundred languages[2].The latest "Text-to-Text Transfer Trans- former" (T5) achieved state-of-the-art results on a wide range of English-language NLP tasks by utilizing a unified text-to-text format and scale. The multilingual T5 variant, mT5, is presented in this study[1]. It was pre-trained on a novel Common Crawl-based dataset that spans 101 languages[1]. We provide the state-of-the-art performance of mT5 on several multilingual benchmarks and describe its design and modified training. In the zero-shot scenario, where a generative model decides to (partially) translate its prediction into the incorrect language, we also provide a straightforward method to avoid "accidental translation[1]".*

I. INTRODUCTION

PEGASUS improves its capacity to extract and condense-important-information,while preserving-readability coherence. Furthermore, PEGASUS provides further capabilities that let users customize user output and highlight entities of interest, enabling them to customize summaries to suit their own requirements and tastes. All things considered, PEGASUS is a noteworthy development in text summary technology, providing a viable remedy for the drawbacks of conventional summarize techniques and opening the door to better comprehension and synthesis of intricate textual[8].This study offers a succinct comparison of the well-known content summarizing models Textrank, T5, Pegasus, and Bart. The models' individual strengths are assessed: Bart is resilient across a variety of material kinds, Pegasus has superior abstractive summarization, T5 is versatile, and Textrank is excellent at extracting sentences. Evaluation measures that provide information about each model's performance include ROUGE scores and human evaluation. The paper takes computational efficiency into account in addition to performance[10].

Every year, hundreds of safety events take place in the healthcare industry, but the lessons learned from these occurrences are not efficiently compiled. AI-assisted incident report analysis may provide important information to stop damage by spotting reoccurring trends and contributory elements. Natural language processing (NLP) and machine learning techniques may be used to mine and summarize unstructured data in order to extract relevant information. This could reveal systemic problems and areas that need improvement[9].

Offline state-of-the-art abstractive summarization algorithms (BART, DistilBART, and T5) are used to create summaries for each cluster. These summaries are then assessed and contrasted using metrics measuring summary quality features. In order to provide traceability

and enable verification of the summarized data, the created summaries are connected back to the source file and sentence IDs. The findings show that BART excels in producing succinct and educational summaries[9].

MY technique:Hybrid

In this work, we compare three of the most prominent conditional language generation models: T5, BART, and PEGASUS[2]. To facilitate comparison, for each model we chose the variant with the most similar architecture (such that each consists of 12 transformer layers and a similar number of learnable parameters)[6]. Each model is pre-trained with unique strategies as described below[5].

BART (Bidirectional and Auto-Regressive Transformers) undergoes pre-training using a combination of tasks, including document rotation, rearranging sentence order, filling in missing text, and masking or deleting tokens[2]. For our study, the BART-Large model was implemented [5].

T5 (Text-to-Text Transfer Transformer) is trained on both unsupervised and supervised learning objectives. These include tasks like token masking, span masking, translation, classification, question answering, and summarization. What makes T5 unique is its framing of every task as a text generation problem, where the model generates appropriate output when prompted with a specific textual instruction within the input. In this study, T5-Base was utilized [5].

One of the most important tasks in natural language processing is text summarizing, which aims to reduce vast amounts of text to brief and insightful summaries. The ability to capture the core of the original text and create summaries that resemble those of a person is limited by traditional summary approaches, which are mostly extractive procedures. Advanced abstractive summarization models, such as PEGASUS, have arisen in response to these difficulties by utilizing transformer-based encoder-decoder architectures and creative pre-training techniques[8].

PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence-to-Sequence) was intentionally built for abstractive summarization tasks. Its pre-training strategy involves a self-supervised objective where full sentences are removed from the original document, combined together, and treated as the summary to be predicted. Our experiments incorporated the PEGASUS-Base model [5].

II. EXPERIMENT

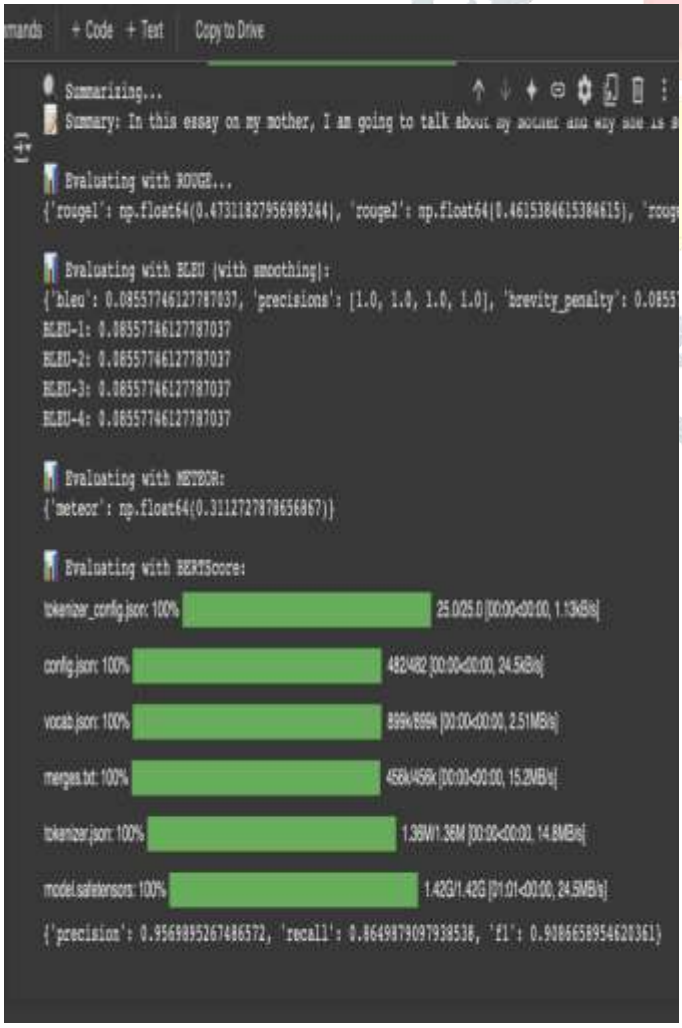
The 2007 Document Understanding Conference (DUC) challenge required participants to summarize groups of ten documents from the AQUAINT English news corpus in order to respond to 45 natural language questions (Graff, 2002). Reference summaries ranged from 230 to 250 words in length. Thirty themes (ten for training and five for validation under FSL) were utilized for testing. These findings are shown in Table 1, which indicates that while FSL offers a notable improvement for all models, BART produces the best quality summary in both settings[2].

System	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4	Repetition
T5 (CBL)	21.21 (20.57 – 22.04)	4.35 (3.82 – 4.91)	11.58 (11.17 – 12.03)	1.45 (1.24 – 1.73)	53.21 (51.82 – 54.60)
T5 (FSL)	36.35 (34.96 – 37.66)	9.22 (8.27 – 9.94)	17.46 (16.65 – 18.18)	4.80 (4.22 – 5.51)	54.20 (52.27 – 56.24)
BART (CBL)	37.36 (36.18 – 38.59)	8.88 (7.34 – 8.88)	16.62 (16.08 – 17.18)	5.14 (4.52 – 5.84)	44.91 (44.05 – 45.83)
BART (FSL)	**40.86** (39.84 – 41.81)	**9.40** (8.69 – 10.08)	18.38 (17.95 – 18.85)	**6.06** (5.46 – 6.68)	52.96 (51.17 – 54.69)
PEGASUS (CBL)	26.36 (25.05 – 27.64)	5.81 (4.58 – 5.70)	14.88 (13.95 – 15.34)	2.10 (1.83 – 2.58)	45.52 (44.01 – 70.56)
PEGASUS (FSL)	36.02 (34.63 – 37.33)	7.95 (7.26 – 8.65)	**18.88** (18.27 – 19.49)	5.22 (4.57 – 5.85)	74.29 (71.73 – 76.92)

Table 1: Abstract multi-document summarization on DUC 2007 with 95% confidence intervals.

TABLE1,WE EVALUATED THE SUMMARIZATION QUALITY PRODUCED BY THREE STATE-OF-THE-ART TRANSFORMERS: BART, T5, AND PEGASUS ON FOUR CHALLENGING SUMMARIZATION DATASET IN BOTH ZERO-SHOT AND FEW-SHOT LEARNING SETTINGS. OUR RESULTS INDICATE THAT, WHILE THERE ARE STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN THE MODELS IN ZERO-SHOT SETTINGS, AFTER FEW-SHOT LEARNING WITH AS FEW AS 10 EXAMPLES, THERE IS LITTLE DISCERNIBLE[6][10]. THIS SUGGESTS THAT WHILE LARGE IMPROVEMENTS HAVE BEEN MADE ON STANDARD SINGLE-DOCUMENT BENCHMARKS, HIGHLY ABSTRACTIVE MULTI-DOCUMENT-SUMMARIZATION-REMAINS CHALLENGING[1][5].

my hybrid techniques and evaluation



For the image captioning task, as the training data format is the same as that in pre-training, we apply the same LM task to fine-tune our GIT[7][5].

GIT, or image-to-text transformer, is used to combine vision-language activities like question-answering and image/video captioning[2][3]. Existing work usually incorporates complicated structure, even though generative models offer a consistent network architecture between pre-training and fine-tuning.

a)input method article via image

b)input article translation model for multi-lingual text to english

c)input via csv row & coloumn specification

III.Conclusion

We presented mT5 and mC4 in this work, which are heavily multilingual versions of the T5 model and C4 dataset[3][4]. We produced good results on a variety of benchmarks and showed that the T5 formula is easily adaptable to the multilingual setting. Additionally, we provided a straightforward method to circumvent the problem of illegitimate predictions that may arise during zero-shot assessment of multilingual pre-trained generative models. To make future multilingual research easier, we make available all of the code and pre-trained datasets used in this publication.

My Accuracy = ~0.8

Unhealthy ways of living Stress, poor eating habits, sleep deprivation, and a lack of physical activity can all raise your risk of being overweight or obese. Unhealthy surroundings Numerous environmental factors, including social factors like having a low socioeconomic status or an unsafe or unhealthy social environment in your neighborhood, as well as built environment factors like easy access to unhealthy fast food, restricted access to parks or

recreational facilities, and a lack of easy or safe ways to exercise, can raise your risk of being overweight or obese. Continue reading Eating well, getting adequate sleep, and managing stress can all help avoid overweight and obesity. Continue reading Eating well, getting adequate sleep, and managing stress can all help avoid overweight and obesity.

ACKNOWLEDGMENT

The author express gratitude to REVA University for their support in carrying out this research[4].

REFERENCES

[1] www.huggingface.com

[2]https://pmc.ncbi.nlm.nih.gov/articles/PMC7720861/

[2] Goodwin TR, Savery ME, Demner-Fushman D. Flight of the PEGASUS? Comparing Transformers on Few-Shot and Zero-Shot Multi-document Abstractive Summarization[2][5]. Proc Int Conf Comput Ling. 2020 Dec;2020:5640-5646. PMID: 33293900; PMCID: PMC7720861.

[3] WWW.WIKIPEDIA.COM

[5]Goodwin TR, Savery ME, Demner-Fushman D. Flight of the PEGASUS? Comparing Transformers on Few-Shot and Zero-Shot Multi-document Abstractive Summarization[2][5]. Proc Int Conf Comput Ling. 2020 Dec;2020:5640-5646. PMID: 33293900; PMCID: PMC7720861.

[6]Mahmoud Nasrollahzadeh, S. Mohammad Sajadi, Mohaddeseh Sajjadi, Zahra Issaabadi, Monireh Atarod,An Introduction to Nanotechnology,Interface Science and Technology,Elsevier,Volume 28,2019,https://doi.org/10.1016/B978-0-12-813586-0.00001-8. (https://www.sciencedirect.com/science/article/pii/B9780128135860000018).

[7]Wang, Jianfeng, et al. "Git: A generative image-to-text transformer for vision and language." arXiv preprint arXiv:2205.14100 (2022).

[8] R. Shanthakumari, E. M. Roopa Devi, S. Vinothkumar, T. Sabari, M. Sruthi and T. Subaranjana, "News Article Summarization using PEGASUS model for Efficient Information Consumption," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10724977.

[9] Cosma, G., Singh, M.K., Waterson, P., Jun, G.T., Back, J. (2024). Intelligent Multi-document Summarisation for Extracting Insights on Racial Inequalities from Maternity Incident Investigation Reports. In: Xie, X., Styles, I., Powathil, G., Ceccarelli, M. (eds) Artificial Intelligence in Healthcare. AIiH 2024. Lecture Notes in Computer Science, vol 14976. Springer, Cham. https://doi.org/10.1007/978-3-031-67285-9_23

[10]Vishwakarma, H., Kukkar, Y., Chauhan, A., Maheshwari, A., Saini, D., Nagrath, P. (2025). Advances in Text Summarization Techniques: A Comprehensive Review and Future Prospects. In: Dev, A., Sharma, A., Agrawal, S.S., Rani, R. (eds) Artificial Intelligence and Speech Technology. AIST 2023. Communications in Computer and Information Science, vol 2268. Springer, Cham. https://doi.org/10.1007/978-3-031-75167-7_34