SURVEY OF LARGE LANGUAGE MODELS ON THE TEXT GENERATION TASK

Michaela Veselá[1], Oldřich Trenz[1]

[1]Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic

ABSTRACT

This paper focuses on the comparison of GPT, GPT-2, XLNet, T5 models on text generation tasks. None of the autoencoder models are included in the comparison ranking due to their unsuitability for text generation tasks. The comparison of the models was performed using the BERT-score metric, which calculates precision, recall and F1 values for each sentence. The median was used to obtain the final results from this metric. A preprocessed dataset of empathetic dialogues was used to test the models, which is presented in this paper and compared with other datasets containing dialogues in English. The tested models were only pre-trained and there was no fine-tune on the dataset used for testing. The transformers library from Hugging face and the Python language were used to test the models. The research showed on the pre-trained dataset empathic dialogues has the highest precision model T5, recall and F1 has the highest precision model GPT-2.

Keywords: natural language processing, auto-regressive transformers, large-scale model, natural language generation, decoder transformer, auto-encoding transformers, sequence to sequence model

JEL Code: C45, L86

1 INTRODUCTION

There has been a lot of progress in the area of large language models in the last few years. According to Tunstall et al. (2022), a major breakthrough in this area is taken to be the development of the Transformers model in 2017, which performed better than the previously used recurrent neural networks (RNNs) and algorithms such as (Stastny and Skorpil, 2007) or hybrid algorithms such as (Stastny et al., 2021), both in terms of machine translation tasks and even in training costs.

Another turning point in the development of large language models was the development of the BERT and GPT models, in 2018 (Tunstall et al., 2022). The BERT model falls into the group of models containing auto-encoding transformers (Devlin et al., 2018) and the GPT model contains auto-regressive transformers (Radford et al., 2018). The group of large language

https://doi.org/10.11118/978-80-7509-990-7-0195

models still includes models containing sequence to sequence transformers. Models containing this transformer are suitable for machine translation. While models containing auto-encoding transformer are suitable for text classification and models containing auto-regressive transformers are suitable for text generation tasks (Rahali and Akhloufi, 2023).

With the passage of time, improvements have been made to both the BERT model and the GPT model. The BERT model has been further modified under the names RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019) or ELECTRA (Clark, 2020). The GPT model has retained its name and resorted to referring to versions of the models using numbers, such as GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023). The GPT model has also come to the attention of the general public due to the fact that OpenAI has made the model freely available to the public, who immediately took a great liking to it.

Currently, the biggest challenge in the field of large-scale language models, and especially for models designed for text generation tasks, is how to perform proper evaluation of the generated language (Thoppilan et al., 2022). The difficulty in evaluating models lies in the fact that for a properly generated language, a large number of aspects need to be addressed, ranging from grammar to coherence, and hence there is no simple way to perform natural language evaluation. One possible way is human annotation, but this is very expensive. Thus, the aim of this paper is to focus on the possible evaluation of language models designed for natural text generation.

2 METHODOLOGY

The methodology section introduces the datasets that contain the dialogues, the metrics used within this paper, and finally the models and their suitability for certain tasks.

2.1 Datasets

To evaluate models in the domain of conversation, datasets containing dialogues can be used. Since the goal is to explore models suitable for implementing a conversational client, datasets containing dialogs need to be selected. Due to the small number of datasets with dialogues in Czech, datasets in English will be used. These include the Ubuntu Dialogue Corpus dataset (Lowe et al, 2015), ProsocialDialog (Kim et al., 2022), SODA (Kim and Hessel, 2022), DailyDialog (Li, 2017) or DialogCC, which also contains images (Lee, 2022). In this case, the empathetic dialogues dataset will be used. The empathetic dialogues dataset contains 76.7 thousand rows of training data, 12 thousand rows of validation data and 10.9 thousand rows of test data (Li et al., 2020) This is a dataset that contains a transformers library designed for the python programming language. The dataset can be used for both text emotion classification and evaluation of generated text. For this reason, it was also selected because it contains emotions.

2.2 Metrics

To evaluate the generated text, it is also possible to use a large number of metrics. To evaluate the generated text, word-based metrics and contextual metrics can be used. Word-based metrics compare the words contained in the generated text and the reference text. While context-based metrics are mostly created as trained models for evaluation and their accuracy has been verified by human annotation (Sai, 2020). A lot of computational power is required when using context-based metrics. One of the possible metrics is BERT-score which is word based. The metric looks for a match using cosine similarity and takes into account the frequency of the document. Its great advantage is that it allows to take into account the importance of words (Zhang et al., 2019). It is a metric that is a good compromise between computational

power and the results it provides, so it will be used in the following paper. In order to use the BERT-score metric, the empathetic dialogues dataset needs to be modified and the reference text needs to be determined.

2.3 Models

The large language models that contain transformers can be divided into three groups, namely Autoregressive models which are suitable for generation tasks, sequence2sequence models which are suitable for machine translation and auto-encoder models which are suitable for text classification (Rahali and Akhloufi, 2023). Therefore, in order to compare the difference between the different types of models, two models that are not designed for text generation were also included in the comparison. These are the BERT model (Devlin, 2018) and the T5 model (Raffel et al., 2019). The models suitable for text generation tasks are XLNet (Yang et al., 2019), GPT (Radford et al., 2018), and GPT-2 (Radford et al., 2019). The models were selected based on the research conducted. Similar to the dataset, the Transformers library from Hugging Face can be used for the models. The library contains already pre-trained models that do not need to be trained further. If the user would still like to retrain the model for a specific type of task, fine-tuned functions are available.

3 RESULTS

Before testing the large language models, it is necessary to prepare the dataset. The empathetic dialogues dataset contains the conversation id (conv_id), utterance id (utterance_idx), what is the context of the dialogue (context), prompt, speaker id (speaker_idx), utterance (utterance), selfeval and tags. A sample dataset is shown in Table 1.

Because the BERT-score metric needs both the generated text and the reference text. It is possible to take every other answer as a reference within a dialogue. It is possible to notice that it is not possible to divide speakers into reference and those whose utterance is marked as the basis for the generated text, for the reason that the same spec may appear in multiple dialogues. Another of the snags in using this dataset can occur if there are an odd number of utterances in a conversation, then it is possible that the model will start generating a response even to an utterance where no response is desired. Should this case occur, it is possible to set the reference response to an empty string. The last situation that could occur is that the model would answer correctly, however it does not match the reference text. Since the Bert-score metric is used, which also takes into account the importance weight of each word, this problem should be avoided by the metric used. Because the crucial words are always repeated in the answers. In the original article, the values obtained from the BERT-score are set as the median (Zhang et al., 2019). This article uses the median to determine precision, recall and f1 median because the BERT-score metric at the point when the dialogue did not continue and thus the person would not respond to the message, but the model generated the text, the metric calculated all the resulting values as zero. So, what happened was that the resulting values contain the extreme outlier of zero, so it is more convenient to use the median.

4 DISCUSSION AND CONCLUSIONS

Testing has shown that T5 has the best results in precision. Recall and F1 has the best GPT-2 though. Sivarajkumar and Wang (2022) also tested GPT-2 and T5 but on classification task and they came out with precision for GPT-2 ranging from 0.73 to 0.50 and recall for T5 ranging from 0.64 to 0.71. 57 to 0.75 and for the T5 ranged from 0.61 to 0.72, for the F1 model the GPT-2 ranged from 0.55 to 0.74 and for the T5 ranged from 0.65 to 0.71. From the above

conv_id	utterance_idx	context	prompt	speaker_idx	utterance	selfeval	tags
hit:1_conv:2	1	afraid	i used to scare for darkness	2	it feels like hitting to blank wall when i see the darkness	4|3|4_3|5|5	
hit:1_conv:2	2	afraid	i used to scare for darkness	3	Oh ya? I don't really see how	4|3|4_3|5|5	
hit:1_conv:2	3	afraid	i used to scare for darkness	2	dont you feel so.. its a wonder	4|3|4_3|5|5	
hit:1_conv:2	4	afraid	i used to scare for darkness	3	I do actually hit blank walls a lot of times but i get by	4|3|4_3|5|5	
hit:1_conv:2	5	afraid	i used to scare for darkness	2	i virtually thought so.. and i used to get sweatings	4|3|4_3|5|5	
hit:1_conv:2	6	afraid	i used to scare for darkness	3	Wait what are sweatings	4|3|4_3|5|5	
hit:1_conv:3	1	proud	I showed a guy how to run a good bead in welding class and he caught on quick.	3	Hi how are you doing today	3|5|5_4|3|4	<HI>
hit:1_conv:3	2	proud	I showed a guy how to run a good bead in welding class and he caught on quick.	2	doing good.. how about you	3|5|5_4|3|4	
hit:1_conv:2	1	afraid	i used to scare for darkness	2	it feels like hitting to blank wall when i see the darkness	4|3|4_3|5|5	

Tab. 1 Empathetic dialogues dataset pattern
Source: Li, Q. et al., 2020

results, it can be seen that Sivarajkumar and Wang (2022) also came out that the T5 model has better precision than GPT-2 in one case, but otherwise it can be generally said that GPT-2 shows better results than T5. Khaliq et al (2022) compared T5 and GPT using BLEU score, where it came out that T5 has better results than GPT. However, this metric is intended for machine translation evaluation (Papineni et al., 2002). A comparison of XLNet was found only with BERT and RoBERTa models, where XLNet showed better results than BERT in most cases. (Liu, 2019) However, a comparison that focused strictly on evaluating the generated text of each model was not found. The research conducted above found that the evaluation results of all the models tested were higher than 62%. It showed that although the T5 model has

Models	precision	recall	f1	Numbers of parameters
XLNet base	0.626486	0.705383	0.665162	110 million
GPT2	0.649675	0.717464	0.684829	1 500 million
T5-base	0.6751	0.682623	0.680336	220 million
GPT	0.633322	0.711081	0.671812	117 million

Tab. 2 Model evaluation results
Source: numbers of parameters: Radford, Alec, et al., 2019; Jangir, S. 2021; Raffel, C. et al., 2019, Nguyen-Mau, T., 2024

significantly fewer parameters than the GPT-2 model, when precision is used, the T5 model performs better than the GPT-2.

Acknowledgements

The publication was supported by the IGA doctoral project (Internal Grant Agency of the PEF) and specifically by the project IGA-PEF-DP-23-023.

This paper was supported by the project CZ.02.1.01/0.0/0.0/16_017/0002334 Research Infrastructure for Young Scientists, this is co-financed from Operational Programme Research, Development and Education.

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

REFERENCES

BROWN, T. B., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A., AGARWAL, S., HERBERT-VOSS, A., KRUEGER, G., HENIGHAN, T., CHILD, R., RAMESH, A., ZIEGLER, D. M., WU, J., WINTER, C. ... and AMODEI, D. 2020. Language Models are Few-Shot Learners (Version 4). *arXiv*. https://doi.org/10.48550/ARXIV.2005.14165

CLARK, K., LUONG, M.-T., LE, Q. V. and MANNING, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2003.10555

DEVLIN, J., CHANG, M.-W., LEE, K. and TOUTANOVA, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Version 2). *arXiv*. https://doi.org/10.48550/ARXIV.1810.04805

JANGIR, S. 2021. *Finetuning BERT and XLNet for Sentiment Analysis of Stock Market Tweets using Mixout and Dropout Regularization*. Technological University Dublin. https://doi.org/10.21427/K0YS-5B82

KHALIQ, Z., FAROOQ, S. U. and KHAN, D. A. 2022. A deep learning-based automated framework for functional User Interface testing. *Information and Software Technology*, 150, 106969. Elsevier BV. https://doi.org/10.1016/j.infsof.2022.106969

KIM, H., HESSEL, J., JIANG, L., WEST, P., LU, X., YU, Y., ZHOU, P., BRAS, R. L., ALIKHANI, M., KIM, G., SAP, M. and CHOI, Y. 2022. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization (Version 3). *arXiv*. https://doi.org/10.48550/ARXIV.2212.10465

KIM, H., YU, Y., JIANG, L., LU, X., KHASHABI, D., KIM, G., CHOI, Y. and SAP, M. 2022. ProsocialDialog: A Prosocial Backbone for Conversational Agents (Version 2). *arXiv*. https://doi.org/10.48550/ARXIV.2205.12688

LAN, Z., CHEN, M., GOODMAN, S., GIMPEL, K., SHARMA, P. and SORICUT, R. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations (Version 6). *arXiv*. https://doi.org/10.48550/ARXIV.1909.11942

LEE, Y.-J., KO, B., KIM, H.-G. and CHOI, H.-J. 2022. DialogCC: Large-Scale Multi-Modal Dialogue Dataset (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2212.04119

LI, Q., LI, P., REN, Z., REN, P. and CHEN, Z. 2020. Knowledge Bridging for Empathetic Dialogue Generation (Version 3). *arXiv*. https://doi.org/10.48550/ARXIV.2009.09708

LI, Y., SU, H., SHEN, X., LI, W., CAO, Z. and NIU, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.1710.03957

LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L. and STOYANOV, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.1907.11692

LOWE, R., POW, N., SERBAN, I. and PINEAU, J. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems (Version 3). *arXiv*. https://doi.org/10.48550/ARXIV.1506.08909

NGUYEN-MAU, T., LE, A.-C., PHAM, D.-H. and HUYNH, V.-N. 2024. An information fusion based approach to context-based fine-tuning of GPT models. *Information Fusion*, 104, 102202. Elsevier BV. https://doi.org/10.1016/j.inffus.2023.102202

OPENAI, ACHIAM, J., ADLER, S., AGARWAL, S., AHMAD, L., AKKAYA, I., ALEMAN, F. L., ALMEIDA, D., ALTENSCHMIDT, J., ALTMAN, S., ANADKAT, S., AVILA, R., BABUSCHKIN, I., BALAJI, S., BALCOM, V., BALTESCU, P., BAO, H., BAVARIAN, M. ... ZOPH, B. 2023. GPT-4 Technical Report (Version 4). *arXiv*. https://doi.org/10.48550/ARXIV.2303.08774

PAPINENI, K. et al. 2002. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

RADFORD, A., NARASIMHAN, K., SALIMANS, T. and SUTSKEVER, I. 2018. *Improving language understanding by generative pre-training*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

RADFORD, A., WU, J., CHILD, R., LUAN, D., AMODEI, D. and SUTSKEVER, I. 2019. *Language Models are Unsupervised Multitask Learners*. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

RAFFEL, C., SHAZEER, N., ROBERTS, A., LEE, K., NARANG, S., MATENA, M., ZHOU, Y., LI, W. and LIU, P. J. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (Version 4). *arXiv*. https://doi.org/10.48550/ARXIV.1910.10683

RAHALI, A. and AKHLOUFI, M. A. 2023. End-to-End Transformer-Based Models in Textual-Based NLP. *AI*, 4(1), 54–110. MDPI AG. https://doi.org/10.3390/ai4010004

SAI, A. B., MOHANKUMAR, A. K. and KHAPRA, M. M. 2020. A Survey of Evaluation Metrics Used for NLG Systems (Version 2). *arXiv*. https://doi.org/10.48550/ARXIV.2008.12009

SIVARAJKUMAR, S. and WANG, Y. 2022. HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2203.05061

STASTNY, J. and SKORPIL, V. 2007. Analysis of Algorithms for Radial Basis Function Neural Network. In: *Personal Wireless Communications*. Springer New York, vol. 245, pp. 54-62, ISSN 1571- 5736, ISBN 978-0-387-74158-1, WOS:000250717300005.

STASTNY, J., SKORPIL. V., BALOGH, Z. and KLEIN, R. 2021. Job shop scheduling problem optimization by means of graph-based algorithm. *Applied Sciences*, 11(4), 1921. ISSN 2076-3417. https://doi.org/10.3390/app11041921

THOPPILAN, R., DE FREITAS, D., HALL, J., SHAZEER, N., KULSHRESHTHA, A., CHENG, H.-T., JIN, A., BOS, T., BAKER, L., DU, Y., LI, Y., LEE, H., ZHENG, H. S., GHAFOURI, A., MENEGALI, M., HUANG, Y., KRIKUN, M., LEPIKHIN, D., QIN, J. ... and LE, Q. 2022. LaMDA: Language Models for Dialog Applications (Version 3). *arXiv*. https://doi.org/10.48550/ARXIV.2201.08239

TUNSTALL, Lewis, WERRA, Lenadro von, WOLF, Thomas and GÉRON, Aurélien. 2022. *Natural language processing with transformers: building language applications with Hugging Face*. Revised edition. Beijing: O'Reilly. ISBN 978-1-098-13679-6

YANG, Z., DAI, Z., YANG, Y., CARBONELL, J., SALAKHUTDINOV, R. and LE, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding (Version 2). *arXiv*. https://doi.org/10.48550/ARXIV.1906.08237

ZHANG, T., KISHORE, V., WU, F., WEINBERGER, K. Q. and ARTZI, Y. 2019. BERTScore: Evaluating Text Generation with BERT (Version 3). *arXiv*. https://doi.org/10.48550/ARXIV.1904.09675

ZIEGLER, D. M., WU, J., WINTER, C., ... and AMODEI, D. 2020. Language Models are Few-Shot Learners (Version 4). *arXiv*. https://doi.org/10.48550/ARXIV.2005.14165

Contact information

Michaela Veselá: email: xvesela@mendelu.cz
Oldřich Trenz: email: oldrich.trenz@mendelu.cz