IRSTI 50.01.11

Guldana Muzdybayeva[1], *Madina Ostemirova*[2], *Dinara Khashimova*[3*]
[1,2,3]SDU University, Kaskelen, Kazakhstan
*e-mail: dinara.khashimova@sdu.edu.kz

RESEARCH OF GPT ALGORITHMS AND ANALYSIS FOR SOME LANGUAGES TO SUGGEST THE BEST WAY TO TRANSLATE INTO KAZAKH

Abstract. The landscape of Natural Language Processing (NLP) has witnessed an expansive array of studies, each tailored to address the unique challenges posed by languages from diverse linguistic back- grounds. This paper offers a thorough summary of relevant publications with a particular focus on language models for the following languages: French, Korean, Russian, Turkish, Chinese, Arabic, Bulgarian, Italian, and Indian (including Hindi and Gujarati). Additionally, the research addresses the challenges and limi- tations involved in the development and application of language models, particularly in Qazaq languages, and provides possible solutions to these problems.

Key words: Natural Language Processing, language models, GPT architectures, Qazaq GPT

1. Introduction

The field of Natural Language Processing (NLP) has seen significant advancements in recent years, with the development of various language models tailored to specific languages [3]. However, many of these models are still primarily designed for English and may not adequately address the unique challenges posed by languages from diverse linguistic backgrounds[9] .This paper aims to provide a comprehensive overview of related works in NLP, focusing on language models for a variety of languages, including French, Korean, Russian, Indian languages (including Hindi and Gujarati), Turkish, Chinese, Arabic, Bulgarian, and Italian. The paper also discusses the challenges and limitations faced in the development and application of language models, particularly in non-English languages, and highlights the importance of adapting and fine-tuning models to address the unique challenges posed by each language. By exploring various deep learning models and techniques applied to different languages, the paper emphasizes the need for continued research and development in NLP to address the challenges and limitations faced in constructing language models for diverse languages [3]. By learning from the experiences and techniques discussed in the related works, future research can

potentially lead to the development of more effective and robust language that can better serve the needs of various languages and applications in the field of NLP.

2. Related work

There has been a wide range of research conducted in the field of natural language processing (NLP), all aimed at tackling the particular difficulties presented by languages with different linguistic backgrounds. This section offers a thorough summary of relevant works, arranged according to the particular emphasis on language models for the following languages: French, Korean, Russian, Indian (including Hindi and Gujarati), Turkish, Chinese, Arabic, Bulgarian, and Italian.

French Language Models: The work [7] is crucial in the effort to create strong French language models. This work emphasizes the necessity for broader models to address the intricacies of the language while navigating the difficulties of dealing with French. The authors examine the drawbacks of current models, like GPTfr and BARThez, and suggest PAGnol-XL as an alternative for more difficult assignments.

Korean Language Models: The research landscape for Korean language understanding is enriched by the in- troduction of KoreALBERT [8]. This study highlights the need of addressing language-specific structures and subtleties by pretraining a lite BERT model for Korean. The authors provide a thorough examination of Ko- rean language model pretraining using metrics such as Word Order Prediction (WOP), Masked Language Model (MLM), and Sentence Order Prediction (SOP).

Russian Language Models: In the quest for enhancing contextually coherent narrow-profile text generation in Russian, the study titled [10] emerges. This study aims to train models specifically for Russian, addressing the lack of subject-specific text production in the language and acknowledging the shortcomings of existing models that mostly serve the English language.

Indian Language Models: The work [11] provides an in-depth examination of the linguistic diversity of Indian languages. The research, preparation techniques, and resource scarcity for languages like Gujarati and Hindi are addressed in this paper. By optimizing models such as PEGASUS, BRIO, and T5, the research advances natural language processing (NLP) solutions for the Indian language environment.

Turkish Language Models: The paper [2] discusses the difficulties in creating efficient Turkish language models.

This study overcomes the structural disparities between Turkish and English by customizing a QA system for the banking industry using Bidirectional Encoder Representations from Transformers (BERT).

Chinese Language Models: In the research paper [6], the investigation of matching-based models for responding to Chinese questions is crucial. By addressing issues including word mismatch, varying expressions, and variations in syntactic structures, this work improves brief text matching in Chinese by introducing a lattice-based CNN model.

Arabic Language Models: The work [1] examines the subtleties of producing meaningful verses in the context of Arabic poem generation. The study uses BLEU ratings for evaluation, emphasizing the difficulties in preserving meaning and coherence and placing special emphasis on the performance of the Markov-LSTM model.

Bulgarian Language Models: One of the main areas of research for [12] is the identification of textual deepfakes in Bulgarian on social media. The study highlights the difficulties brought about by Bulgarian's low resource availability, such as poor machine translation and a dearth of appropriate detectors for identifying textual deep- fakes in the language.

Italian Language Models: The article "As Good as New." The translation of English GPT-2 into Italian and Dutch is examined in [13]. With a focus on closest neighbor preservation, embedding similarity, and perplexity, this work shows that lexical embeddings can be transformed for efficient text creation in Italian.

In summary, the wide range of relevant research that are showcased here highlights the coordinated efforts to customize NLP models to the nuances of different languages, which helps to create a more effective and inclusive global NLP environment.

3. Dataset

This section provides a thorough exploration of data collection methodologies across various studies related to language models, encompassing GPT architectures, Arabic poetry generation[1] [4], Chinese question answering[6], Turkish question answering [2], Indian language summarization [11], Korean language understanding[8], Bul- garian text generation[12], Italian language adaptation[13].

3.1. Data Collection for GPT Architectures

Key Concepts in Data Collection for GPT Architectures.

Data collection for GPT architectures involves critical concepts:

Large and Diverse Text Corpus:

- Gather diverse corpora for contextual understanding across topics, styles, and genres.

Preprocessing:

- Tokenization, lowercasing, and special character removal prepare the corpus for training.

Training Data Preparation:

- Split the preprocessed corpus into training, validation, and test sets for effective model training.

Relevance:

- Ensure high-quality, relevant data for enhanced model performance and generalization.

Ethical Considerations:

- Adhere to privacy and copyright laws, minimize biases, and ensure diverse representation.

Licensing and Permissions:

- Choose openly available, properly licensed data sources to ensure legal compliance.

Case Studies in Data Collection for GPT Architectures

- **Russian Text Generation:**

· Utilized articles from ”Populyarnaya Mekhanika” and ”ForkLog” portals.

· **Arabic Poetry Generation:[1] [4]**

· Pretrained GPT-2 using Khaleej-2004 and Watan-2004 corpora, fine-tuned on an Arabic poetry dataset.

· **Chinese Question Answering:[6]**

· Utilized NLPCC-2016 evaluation task datasets, addressing challenges in short text matching.

· **Turkish Question Answering:[2]**

Employed datasets including Wikipedia Corpus (Tr), News Corpus (Tr), and Banking Sector QA (Tr).

· **Indian Language Summarization:[11]**

· Used the ILSUM 2022 dataset for English, Hindi, and Gujarati, fine-tuning models like PEGASUS and BRIO.

· **Korean Language Understanding:[8]**

· Pretraining corpora from web news, Korean Wikipedia, NamuWiki, and Book Corpus. Applied Masked Language Modeling, Sentence Order Prediction, and Word Order Prediction.

Data Collection for Turkish Question Answering System

The study on [2] conducted research using Turkish datasets, including the Wikipedia Corpus (Tr), News Corpus Research of GPT algorithms and analysis for some languages to suggest the best way to translate into Kazakh

(Tr), Economy Corpus (Tr), SQuAD (Tr), NewsQA (Tr), and Banking Sector QA (Tr). The evaluation of the system's performance was carried out in both open and closed domains, reflecting the comprehensive assessment of its capabilities across a diverse range of datasets and domains in the Turkish language.

Data Collection for Chinese Question Answering

The investigation into [6] centered on two primary objectives. Firstly, the study leveraged datasets from the NLPCC-2016 evaluation task, emphasizing the utilization of relevant and specific data for Chinese question answering. Secondly, the research aimed to tackle challenges associated with short text matching in the Chinese language. To address these challenges effectively, the study proposed and implemented efficient Lattice CNNs, emphasizing their practical scalability. The incorporation of Lattice CNNs was designed not only to enhance the model's matching capabilities but also to ensure its feasibility and efficiency in real-world applications, aligning with the complexities and nuances of Chinese language processing.

Data Collection for Korean ALBERT Model

The paper on [8] detailed:

· Pretraining corpora from web news articles, Korean Wikipedia, NamuWiki, and book corpus . The web news articles were collected from eight major Korean newspapers across various topics, including politics, social, economics, culture, IT, opinion, and sports, from January 1, 2007, to December 31, 2019. The Korean Wikipedia and Namu Wiki were crawled in October 2019 and December 2019, respectively. The book corpus included plots and editorial reviews about all Korean books published in 2010 to December 31, 2019.

· Text preprocessing techniques, including removal all meta-tags such as the date of writing and name(s) of the author(s) in newspapers.

· Training objectives involving Masked Language Modeling, Sentence Order Prediction, and Word Order Prediction.

Data Collection for Indian Language Summarization

The paper on [11] provided insights into:

· Creating the ILSUM 2022 dataset from news articles in Indian English, Hindi, and Gujarati.

· Data preprocessing steps, including removal of punctuations and delimiters.

· Fine-tuning pre-trained models for text summarization.

Data Collection for Bulgarian Text Generation. The paper [12] explored three approaches, including creating a new Bulgarian-language dataset. The paper investigates Bulgarian textual deepfake detection via machine translation, LM generation, and classifier training. It combines English datasets with Bulgarian content, tests LM-generated datasets, and proposes real-world applications, highlighting the efficacy of LM detection and manual fact-checking while acknowledging machine translation limitations in this context.

Data Collection for Italian Language Adaptation. The article on [13] proposed adapting GPT-2 models to Italian, using a combination of Wikipedia data and web texts. The paper proposes a method to adapt existing pre-trained English GPT-2 models to Italian and Dutch by retraining lexical embeddings without tuning the Transformer layers. The first step is to retrain the lexical embeddings of the GPT-2 small model without touching the Transformer layers. The retrained lexical embeddings are well aligned with the English vocabulary, and GPT-2 can generate realistic text in Italian and Dutch after this step. The next step is to transform the small lexical embeddings to the GPT-2 medium lexical embedding space. The least-squares regression method is the most effective transformation method for this scaling procedure. The paper also discusses the resources, models, and computation used in the training process.

Data Collection for Arabic Text Summarization. The data collection for GPT-2 architectures involves fine-tuning the model on a specific dataset. In the context of a study [1] , [4] on Arabic poem generation, the researchers collected a dataset of 34,466 verses from praise poems written by different Muslim poets. They pre-processed the data using weights that had previously been trained in English, the authors began training the GPT-2 model in Arabic. Using the Tokenizers Library (Hugging Face), they trained an able tokenizer on an Arabic corpus and obtained the vocabulary files (vocabulary size 50K tokens) of the GPT-2 tokenizer in Arabic. The study used the small version of the GPT-2 model, which was fine-tuned on the dataset using the gpt-2-simple package in Python. The input data to the model is a single text file, and the model was trained to generate praise poems. They calculated the BLEU-1, BLEU-2, BLEU-3 and BLEU-4 scores, and compared the results with the work of (Talafha and Rekabdar, 2019b) and (Talafha and Rekabdar, 2019a). The GPT-2 model better than other models for BLEU-1, BLEU-2, BLEU-3 and BLEU-4

4. Metrics

Metrics 1

Table 1: Evaluation Metrics for Different Languages

Language	Metric	Models	Dataset
Turkish	EM, F-Score	Deep Learning Neural	Corpus,

		Networks	squad,newsqa
French	Model Size, EM, F1, R-1, R-2, R-L	Extra-Large Generative Model	Various NLP Tasks
Chinese	P1, MAP, MRR	Lattice cnns for Matching	DBQA Dataset
Korean	WOP, MLM, SOP	Korealbert	Korean Wikipedia, Book Corpus
Russian	Not specified	GPT-2	Populyarnaya mekhanika, Forklog
Arabic	BLEU (Bi, Tri, 4-gram), Human Evaluation	LSTM, Markov-LSTM, GPT-2	Praise Poems
Indian	ROUGE-1, ROUGE-2, ROUGE-4	Deep Learning-Based Approaches	ILSUM 2022
Bulgarian	Accuracy, Precision, Recall, F1	Textual Deepfakes Detection	Tweepfake3
Italian, Dutch	Embedding Similarity, Perplexity,	GPT-2 Transformation	Itwac corpus, Dutch news websites

Table 2: Experimental results for Arabic Language

Datasets	**Models**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
Khaleej-2004	Vanilla	0.0211	0.0199	0	0
Watan-2004	LSTM	0.1522	0.1124	0.0081	0.0013
Arabic poetry	GRU	0.1512	0.1139	0.0084	0.0021
-	RNN EncoderDecoder (without attention)	0.2513	0.1539	0.0740	0.0510
-	RNN EncoderDecoder (with attention)	0.3010	0.2110	0.0911	0.0801
-	(Talafha and Rekabdar, 2019b) Model	0.4122	0.3144	0.204	0.1092
-	(Talafha and Rekabdar, 2019a) Model	0.5301	0.4010	0.3001	0.1500
-	GPT-2	0.8739	0.5369	0.3230	0.1871

Table 3: Experimental results for Korean Language

Dataset	*Model*	*KorNLI*	*KorSTS spearman*	*NSMC*	*PD*	*NER*	*KorQuAD1.0 f1*	*Avg.*
Korean Wikipedia	*Multilingual BERT*	*76.8*	*77.8*	*87.5*	*91.1*	*80.3*	*86.5*	*83.3*
Book corpus	*XLM-R*	*80.0*	*79.4*	*90.1*	*92.6*	*83.9*	*92.3*	*86.4*
NamuWiki, Web News	*KoBERT*	*78.3*	*79.2*	*90.1*	*91.1*	*82.1*	*90.3*	*85.2*
	ETRI BERT	*79.5*	*80.5*	*88.8*	*93.9*	*82.5*	*94.1*	*86.6*
-	*KoreALBERT Base*	*79.7*	*81.2*	*89.6*	*93.8*	*82.3*	*92.6*	*86.5*

-	*KoreALBERT Large*	*81.1*	*82.1*	*89.7*	*94.1*	*83.7*	*94.5*	*87.5*

Table 4: Experimental results for Turkish Language

Data sets	*Model*	*EM*	*F-Score*
*SQuAD (Tr) and NewsQA (Tr)**	*BERTurk*	*57.43*	*69.36*
	Current Study	*55.26*	*67.07*
	mBERT	*54.52*	*65.74*
*Banking Sector QA (Tr)***	*BERTurk*	*55.89*	*80.87*
	Current Study	*54.09*	*79.01*
	mBERT	*50.74*	*77.03*

Table 5: Experimental results for Indian Language

Dataset	Model	ROUGE-1	ROUGE-2	ROUGE-4
ILSUM 2022	Deep Learning-Based Approaches	0.5559	0.4547	0.4136
ILSUM 2022	Deep Learning-Based Approaches	0.2087	0.1192	0.0838

Table 6: Experimental results for Bulgarian Language

Dataset	Model	Acc.	Precision	Recall	F1
cGPT	BdB		0.93	0.94	0.93
BwB			0.94	0.95	0.95
Human			0.95	0.94	0.95
Total		0.94	0.94	0.94	0.94
cGPT	SVM		0.88	0.82	0.85
BwB			0.89	0.83	0.86
Human			0.87	0.92	0.89
Total		0.88	0.88	0.88	0.87
cGPT	Logistic		0.80	0.80	0.80
BwB	regression		0.85	0.85	0.85

Human			0.87	0.87	0.87
Total		0.85	0.85	0.85	0.85

Table 7: Experimental Results for Italian and Dutch Languages

Dataset	Model	Socia l	Nexxz	Legal	Bxaceedings
Italian data: Wikipedia data (2.8GB), Web texts from the ItWGa corpus	Small model with relearned lexical embeddings	134.6 4	son 67.1 4	16.95	
Dutch data: Wikipedia (2.0GB), newspaper articles		-	239.14	52.01	44.47
Italian data: Wikipedia data (2.8GB), Web texts from the ItWaG corpus	Small model with additional finetuning to the target language	118.1 9	55.63	15.36	
Dutch data: Wikipedia (2.0GB), newspaper articles		-	171.83	42.92	36.35
Italian data: Wikipedia data (2.8GB), Web texts from the ItWaC corpus	Medium model with releamed lexical embeddings	123.6 4	59.18	14.95	
Dutch data: Wikipedia (2.0GB), newspaper articles			234.52	45.01	40.62
Italian data: Wikipedia data (2.8GB), Web texts from the ItWhaC corpus	GPT-2 small based GePpeTto model	179.4 7	80.83	34.71	

5. Methodology

5.1. Deep Learning-Based Approaches for Indian Language Summarization [11]

In addressing the challenges associated with Indian language summarization, the proposed methods draw in- sights from diverse deep learning models. The ILSUM 2022 dataset, encompassing Indian English, Hindi, and Gujarati articles,

forms the basis for experimentation. To handle the raw and diverse nature of the datasets, efficient data cleaning techniques are employed. The average length of articles and extractive summaries guides model architecture decisions. The fine-tuning process involves leveraging pre-trained models such as PEGA- SUS, BRIO, T5, IndicBART, XL-Sum, and mBART. Evaluation metrics, including ROUGE-1, ROUGE-2, and ROUGE-4, are employed to assess the performance of these models on the ILSUM 2022 datasets. The scarcity of resources for certain Indian languages, notably Gujarati, is acknowledged, and the proposed methodology seeks to bridge this gap through innovative adaptations and fine-tuning.

5.2. **KoreALBERT: Pretraining a Lite BERT Model for Korean Language Understanding [8]**

KoreALBERT is a multi-layer bidirectional Transformer encoder with the same factorized embedding parame- terization and cross-layer sharing as ALBERT. Inheriting ALBERT-base, KoreALBERT-base has 12 parameter sharing layers with an embedding size of 128 dimensions, 768 hidden units, 12 heads, and GELU nonlinearities [5]. The total number of parameters in KoreALBERT-base is 12 millions, and it increases to 18-million param- eters for KoreALBERT-large having 1024 hidden dimensions.[8]

5.3. **Turkish Question Answering System Utilizing BERT Algorithm [2]**

For the development of a question answering system in the Turkish language, the proposed methods incorpo- rate the Bidirectional Encoder Representations from Transformers (BERT) algorithm. The system undergoes a two-step process, starting with the generation of a language model using BERT, followed by fine-tuning for question answering (QA) tasks specific to the banking sector. Training datasets, including the Stanford Question Answer Data Set (SQuAD) structure, facilitate the model's learning. Evaluation metrics such as Exact Match (EM) and F-Score gauge the system's performance in responding to factoid questions within the banking domain.

5.4. **Arabic Poem Generation using Deep Learning Models [1][4]**

In the realm of Arabic poem generation, the proposed methods encompass character-based LSTM, Markov- LSTM, and pre-trained GPT-2 models. Trained on a dataset of Arabic praise poems, the models aim to generate coherent verses. Evaluation utilizes BLEU scores, with an emphasis on capturing meaning and coherence. The paper highlights the success of the Markov-LSTM model, surpassing both LSTM and GPT-2 in terms of meaning. The input data, consisting of preprocessed Arabic poem datasets, forms the foundation for the generation of culturally resonant verses.

5.5. **Lattice-Based CNN Model for Chinese Question Answering [6]**

To enhance the matching process in Chinese question answering, the proposed method introduces a novel lattice-based CNN (LCN) model. Leveraging word

lattices as input, the model employs siamese architecture and pooling mechanisms to merge feature vectors for effective matching scores. Evaluation metrics, including Precision1 (P1), Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR), assess the model's per- formance on datasets such as DBQA and KBRE. The emphasis is on addressing challenges related to short text matching in Chinese.

5.6. **Adapting English GPT-2 to Italian and Dutch [13]**

In adapting English GPT-2 models to Italian and Dutch, the proposed methods center around the transfor- mation of lexical embeddings and text generation. Retraining lexical embeddings without tuning Transformer layers facilitates the generation of realistic text in new languages. Metrics such as embedding similarity, per- plexity, and nearest neighbor preservation guide the evaluation process. The paper underscores the adaptability of GPT-2 to new languages and the generation of contextually coherent text.

5.7. **Deep Learning Approaches for Article Summarization in Indian Languages [11]**

The paper consolidates various deep-learning approaches applied to the ILSUM 2022 Indic language summariza- tion datasets. Pre-trained models such as PEGASUS, BRIO, T5, IndicBART, XL-Sum, and mBART undergo fine-tuning with English, Hindi, and Gujarati datasets. ROUGE-1, ROUGE-2, and ROUGE-4 metrics evaluate the effectiveness of the summarization models. Challenges, including dataset resource shortages and prepro- cessing methodologies, are addressed by innovative adaptations and fine-tuning. The proposed methods aim to enhance article summarization in Indian languages through a combination of pre-trained models and tailored pipelines.

5.8. **Deep Learning Approaches for Text Generation in French [7]**

This comprehensive paper explores various deep learning methods applied to text generation tasks specifically tailored for the French language. Notable models, including GPTfr and BARThez, are considered for pre- training, and adaptation techniques are employed to enhance their performance. Metrics such as model size, EM, F1 score, R-1, R-2, and R-L (ROUGE scores) are utilized to evaluate the models across diverse tasks, including classification, paraphrasing, natural language inference, question answering, and summarization. The proposed methods focus on addressing challenges related to dataset pre-processing, model size, and the scarcity of large-scale French language models.

5.9. **Deep Learning Approaches for Text Generation in Russian [10]**

This paper delves into deep learning approaches for text generation in the Russian language, emphasizing the need for contextually coherent narrow-profile text. The study involves training models to generate coherent articles within specific subject areas in Russian. Metrics assessing the quality of generated text, along with a comparison against existing models such as BERT

and ELMo, are employed. The proposed methods aim to address the existing gap in models capable of generating meaningful and contextually coherent text in Russian.

5.10. **Detecting Textual Deepfakes in Bulgarian on Social Media [12]**

This paper focuses on the detection of textual deepfakes in Bulgarian on social media platforms. Challenges associated with low-resource languages, including Bulgarian, are addressed. The proposed methods involve ex- periments with machine translation, training classifiers on Bulgarian datasets, and combining language model text detection with fact-checking. Metrics such as accuracy, precision, recall, and F1 score are used to evaluate the performance of the proposed approaches in identifying textual deepfakes in Bulgarian.

6. Challenges and Solutions in Neural Network Architectures

In this section, we explore various proposed neural network architectures for different natural languages, with a focus on the challenges and assumptions involved in constructing a Qazaq GPT architecture. Each subsection addresses the specific problems encountered in existing literature pertaining to neural network models for different languages, emphasizing the distinctive issues faced by each linguistic context.

Turkish Language: Challenges in Question Answering Systems

The paper [2] identifies multiple challenges in the realm of Turkish language processing. These challenges include the limited success of existing QA systems, difficulties in generating a Turkish language model, and the unsatisfactory performance of BERT in Turkish. Furthermore, the paper emphasizes the need for appropriate methods and guidelines for tasks in languages structurally different from English.

Chinese Language: Lattice CNNs for Matching in Question Answering

In study [6] delves into challenges specific to short text matching in Chinese QA tasks. The identified challenges involve word segmentation issues, diverse expressions leading to relational matching difficulties, and significant differences in styles or syntactic structures. The proposed lattice-based CNN model aims to address these challenges by leveraging multi-granularity information in the word lattice, enhancing its ability to handle noisy information in Chinese question answering.

Korean Language: Pretraining Lite BERT Model (KoreALBERT)

The lack of pretrained ALBERT models specifically designed for the Korean language is addressed in the paper [8]. The main issue is that there are not any specialist models for Korean, so pretraining lite BERT models is necessary to improve language understanding.

French Language: Challenges in Developing Extra-Large Generative Models

The paper [7] outlines challenges in French language model development. Issues include inadequate dataset pre-processing leading to low-

quality outputs, limited availability of large-scale French models, challenges in managing the computational budget for training, and the critical choice of suitable datasets. The authors stress the importance of overcoming these challenges to develop and enhance large-scale French language models.

Russian Language: Contextually Coherent Text Generation

The paper [10] focuses on the absence of models capable of generating contextually coherent narrow-profile text in Russian. The identified problem is the lack of Russian-specific models, with existing solutions mainly dedicated to English. The study aims to fill this gap by training a model for generating coherent articles in Russian for specific subject areas.

Bulgarian Language: Detection of Textual Deepfakes on Social Media

Authors in [12] addresses the challenges in detecting textual deepfakes in Bulgarian due to the language's low- resource nature. The primary problems include the lack of suitable datasets, detectors, and machine-learning techniques for Bulgarian, as well as difficulties arising from unsatisfactory machine translation and low-resource language characteristics. The paper proposes approaches to detect LM-generated texts, emphasizing the need for appropriate datasets, detectors, and techniques.

Arabic Language: Poem Generation Using GPT-2

The paper [1] discusses the successful implementation of a GPT-2 model for generating Arabic poems. However, the challenges in Arabic poem generation include inconsistencies in generated verses, low performance in terms of meaning, difficulty in outperforming baselines, and limited research in the Arabic poem generation domain.

Adapting GPT-2 Models: Italian and Dutch Languages:

In [13] proposes a method for adapting GPT-2 models to Italian and Dutch. While the article does not discuss major problems for these languages, it highlights limitations in training language models for less-resourced languages due to data and computational constraints. The primary focus is on the proposed adaptation method rather than specific language-related challenges.

Indian Languages: Deep Learning-Based Approaches for Article Summarization

The paper [11] outlines challenges in Indian language GPT models. Issues include limited research and resources for low-resource Indian languages, such as Hindi and Gujarati, and challenges in fine-tuning models for specific languages. The paper emphasizes the shortage of relevant datasets and preprocessing methodologies, especially for languages like Gujarati, compared to high-resource languages like English.

Conclusion

In summary, the papers covered shed light on a number of difficulties and constraints encountered during the creation and use of language models, particularly for non-English languages. These hurdles include problems with the

structure and properties of particular languages, the availability of sufficient pretraining datasets, and the requirement for customized models to handle linguistic subtleties. Building a Qazaq GPT architecture may now present comparable difficulties. Possible restrictions and issues with Qazaq GPT could be as follows:

- Limited Pretraining Data: The model's performance and generalization may be impacted by the lack of extensive pretraining data in the Kazakh language.
- Nuances in Kazakh Language: The distinctive linguistic features of the Kazakh language could make it difficult to create models that faithfully represent its subtleties, which could have an effect on the caliber of material that is produced.
- Fine-Tuning Challenges: If pre-existing models are modified, fine-tuning for QazaqGPT might entail diffi- culties akin to those encountered in the Indian language GPT models as well as careful consideration of linguistic quirks.
- Resource Constraints: Achieving the ideal model size and performance may be hampered by limitations in the computational resources available for training large-scale models.
- Evaluation Data and Metrics: Determining how well the model generates contextually coherent and mean- ingful material in Kazakh may depend on the availability of relevant evaluation data and metrics that are specific to Qazaq GPT.

A number of assumptions and strategies can be taken into consideration in order to overcome the potential diffi- culties in building the Qazaq GPT architecture. These assumptions come from standard procedures in machine learning and natural language processing. The following assumptions might direct towards the development of Qazaq GPT::

1. Collaborating Together with Linguistic Specialists:

- Assumption: Working with language experts who have expertise in Kazakh will give you important insights into the nuances, syntax, and expressions of the language.
- Approach: In order to ensure linguistic accuracy, involve linguists in the processes of dataset curation, model training, and fine-tuning.
- Collecting Data via Crowdsourcing:
- Assumption: Using crowdsourcing to involve the Kazakh-speaking community can aid in the collection of authentic and varied linguistic data.
- Approach: Put in place systems for gathering data through crowdsourcing, encouraging people to submit sentences, phrases, and examples of context in Kazakh.
- Integration of Multimodal Data:
- Assumption: Adding multimodal data, such as text, pictures, and maybe audio, can help the model perform better and better understand context.
- Approach: Investigate how to combine various data kinds to develop a more thorough comprehension of the Kazakh language.
- Learning Transfer from Related Linguistics:

- Assumption: QazaqGPT may have a basis thanks to the utilization of pre-existing models trained on comparable Turkish languages.
- Approach: Examine whether knowledge from models trained on languages comparable to Kazakh can be applied to other languages, making necessary adjustments and fine-tuning
- Collaboration with Open Source:
- Assumption: Open source cooperation encourages creativity and makes it possible for a larger community to assist in improving QazaqGPT.
- Strategy: Encourage cooperative development and contributions by making the model architecture, datasets, and resources available to the open-source community.
- The creation of Qazaq GPT can be approached with a comprehensive awareness of the linguistic, cultural, and technological characteristics specific to the Kazakh language by assuming these things and putting related tactics into practice.

References

1 Beheitt, M.E.G., Hmida, M.B.H.: Automatic arabic poem generation with gpt-2. In: ICAART (2). pp. 366–374 (2022)

2 Gemirter, C.B., Goularas, D.: A turkish question answering system based on deep learning neural networks. Journal of Intelligent Systems: Theory and Applications 4(2), 65–75 (2021)

3 Grishman, R., Sundheim, B.M.: Message understanding conference-6: A brief history. In: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics (1996)

4 Hakami, A., Alqarni, R., Almutairi, M., Alhothali, A.: Arabic poems generation using lstm, markov-lstm and pre- trained gpt-2 models. Computer Science & Information Technology (CS & IT) 11, 139–147 (2021)

5 Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)

6 Lai, Y., Feng, Y., Yu, X., Wang, Z., Xu, K., Zhao, D.: Lattice cnns for matching based chinese question answering. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 6634–6641 (2019)

7 Launay, J., Tommasone, E., Pannier, B., Boniface, F., Chatelain, A., Cappelli, A., Poli, I., Seddah, D.: Pagnol: An extra-large french

generative model. arXiv preprint arXiv:2110.08554 (2021)

8 Lee, H., Yoon, J., Hwang, B., Joe, S., Min, S., Gwon, Y.: Korealbert: Pretraining a lite bert model for korean language understanding. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 5551–5557. IEEE (2021)

9 Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007)

10 Shatalov, O., Ryabova, N.: Towards russian text generation problem using openai's gpt-2. CEUR Workshop Pro- ceedings (2021)

11 Tangsali, R., Pingle, A., Vyawahare, A., Joshi, I., Joshi, R.: Implementing deep learning-based approaches for article summarization in indian languages. arXiv preprint arXiv:2212.05702 (2022)

12 Temnikova, I., Marinova, I., Gargova, S., Margova, R., Koychev, I.: Looking for traces of textual deepfakes in bulgarian on social media. In: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. pp. 1151–1161 (2023)

13 de Vries, W., Nissim, M.: As good as new. how to successfully recycle english gpt-2 to make models for other languages. arXiv preprint arXiv:2012.05628 (2020)

Гүлдана Мұздыбаева[1],Мадина Өстемирова[2], Динара Хашимова[3]*
[1,2,3]SDU University, Қаскелең, Қазақстан
*e-mail: dinara.khashimova@sdu.edu.kz

GPT АЛГОРИТМДЕРІН ЗЕРТТЕУ ЖӘНЕ ҚАЗАҚ ТІЛІНЕ АУДАРУДЫҢ ЕҢ ЖАҚСЫ ӘДІСІН ҰСЫНУ ҮШІН КЕЙБІР ТІЛДЕРГЕ ТАЛДАУ ЖАСАУ

Аңдатпа. Табиғи тілді өңдеу (NLP) саласында көптеген зерттеулер жүргізілді, олардың әрқайсысы әртүрлі лингвистикалық шығу тегі бар тілдерді қолданудан туындайтын бірегей мәселелерді шешуге бағытталған. Бұл мақалада француз, корей, орыс, түрік, қытай, араб, болгар, итальян және үнді (соның ішінде хинди және Гуджарати) тілдерінің модельдеріне ерекше назар аудара отырып, тиісті басылымдардың егжей-тегжейлі шолуы берілген. Сонымен қатар, зерттеу

тілдік модельдерді, әсіресе қазақ тілінде әзірлеуге және қолдануға байланысты проблемалар мен шектеулерді қарастырады және осы мәселелердің ықтимал шешімдерін ұсынады.

Түйін сөздер: табиғи тілді өңдеу, тіл модельдері, GPT архитекуралы, Qazaq GPT.

Гулдана Муздыбаева[1],Мадина Остемирова[2], Динара Хашимова[3]*
[1,2,3]SDU University, Каскелен, Казахстан
*e-mail: dinara.khashimova@sdu.edu.kz

ИССЛЕДОВАНИЕ АЛГОРИТМОВ GPT И ИХ АНАЛИЗ ДЛЯ НЕКОТОРЫХ ЯЗЫКОВ, ЧТОБЫ ПРЕДЛОЖИТЬ НАИЛУЧШИЙ СПОСОБ ПЕРЕВОДА НА КАЗАХСКИЙ

Аннотация. В области обработки естественного языка (НЛП) было проведено множество исследований, каждое из которых было направлено на решение уникальных проблем, возникающих в результате использования языков разного лингвистического происхождения. В этой статье представлен подробный обзор соответствующих публикаций с особым упором на языковые модели для следующих языков: французского, корейского, русского, турецкого, китайского, арабского, болгарского, итальянского и индийского (включая хинди и гуджарати). Кроме того, исследование рассматривает проблемы и ограничения, связанные с разработкой и применением языковых моделей, особенно на казахском языке, и предлагает возможные решения этих проблем.

Ключевые слова: обработка естественного языка, язык модели, архитектуры GPT, Qazaq GPT.

Received 14 April 2024