Language Models for Multi-lingual Tasks - a Survey

Amir Reza Jafari, Behnam Heidary, Reza Farahbakhsh, Mostafa Salehi, Noel Crespi

Language Models for Multi-lingual Tasks - a Survey

Amir Reza Jafari*§, Behnam Heidary†§, Reza Farahbakhsh*, Mostafa Salehi† and Noel Crespi*
* Samovar, Telecom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France
† New Sciences and Technologies, University of Tehran, Tehran, Iran
§ Equal contribution

***Abstract*—These days different online media platforms such as social media provide their users the possibility to exchange and engage in different languages. It is not surprising anymore to see comments from different languages in posts published by international celebrities and figures. In this era, understanding cross-language content and multilingualism in natural language processing (NLP) are crucial, and huge amount of efforts have been dedicated on leverage existing technologies in NLP to tackle this challenging research problem, specially with advances in language analysis and the introduction of large language models. In this survey, we provide a comprehensive overview of the existing literature focusing on the evolution of language models with a focus on multilingual tasks and then we identify potential opportunities for further research in this domain.**

Keywords*—*Language Models; Transfer Learning; BERT, NLP; Multilingual task; Low Resource Languages; LLMs

I. Introduction

The exploration of multilingualism across various Natural Language Processing (NLP) tasks stands as one of the most dynamic and challenging within the academic community. Over the past decade, these discussions have surged to the forefront of both linguistic and computer science arenas, specially by the increasing prevalence of transfer learning techniques in NLP. This endeavor gains significance in light of the pervasive influence of social media and the profound engagement of users worldwide with trending topics. The extensive usage of social media platforms underscores the criticality of developing robust multilingual models capable of understanding and processing diverse linguistic inputs.

Due to the growing attention to multilingual models, there arises a pressing need to comprehensively review their evolution, from inception to maturity. Moreover, it is equally vital to assess the monolingual models, as they provide a foundational benchmark for the advancements in multilingual NLP. This comprehensive approach is essential for gaining a detailed understanding of the complex factors involved in multilingual NLP and for guiding future progress.

In the era dominated by transformers, pre-trained models have emerged as a cornerstone in Natural Language Processing (NLP) due to their ability to harness vast datasets and computational resources for training. By leveraging learned representations and parameters, these models adeptly capture intricate patterns and knowledge embedded within the training data [7]. On one hand, transfer learning, a technique widely employed in various machine learning approaches including domain adaptation and multitask learning, serves as a pivotal solution for transferring essential knowledge across tasks [1], [8].

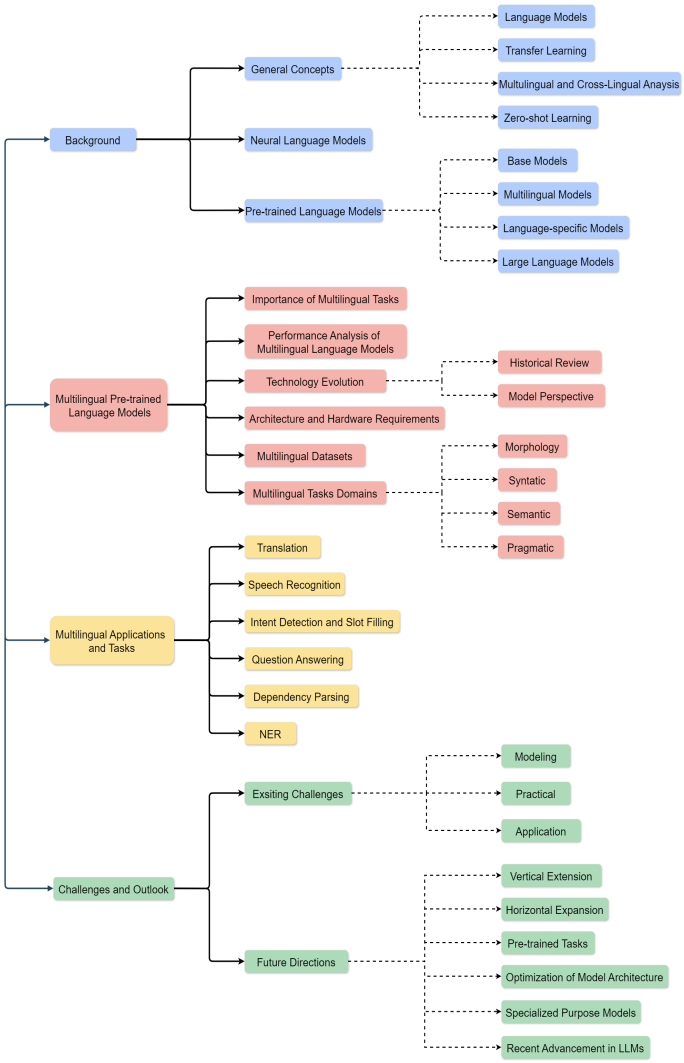

Fig. 1: An overview of the structure of the survey

On the other hand, The concept of multilingualism in language models epitomizes their versatility in understanding and generating text across multiple languages. Trained on diverse datasets encompassing various languages and NLP tasks like machine translation and text processing, these models exhibit the remarkable capability to comprehend and produce text in multiple languages [4], [9]. Our survey endeavors to provide a comprehensive overview of the evolution of language models and the concept of multilingualism across diverse tasks, spotlighting models introduced for languages with lower

TABLE I: Main surveys in the field of language models and multilingual NLP tasks

Title	Main Focus of the study	How differentiate it with our paper
A survey of transfer learning [1]	Mainly focused on transfer learning paradigm and its current solutions and applications applied to transfer learning	An overview of transfer learning with less details and more focus on this paradigm in multilingual models and applications
A survey of cross-lingual word embedding models [2]	provide a comprehensive typology of cross-lingual word embedding models and compare their data requirements and objective functions	we focused on outputs of models and only talk about structure and word embedding enough to help readers to understands outputs
Evolution of transfer learning in natural language processing [3]	This survey provides an comprehensive architectural and technical view of recent advances in transfer learning in models such as BERT, GPT, ELMo, ULMFit.	We focused more on transfer learning in multi-lingual tasks and its evolution timeline instead of a detailed analysis of architectures
A survey of multilingual neural machine translation [4]	This survey presents an in-depth survey of existing literature on MNMT and also categorizes various approaches based on the resource scenarios as well as underlying modeling principles	We have a more general overview of multilingual tasks which include machine translation too but not limited to a specific task
Cross-lingual learning for text processing: A survey [5]	a comprehensive table of all the surveyed papers with various data related to the cross-lingual learning techniques they use	we have a model perspective and focused on multilingual language models more
A Survey on Evaluation of Large Language Models [6]	This survey presents a review of the evolution of large language models and the perspective of related tasks focusing on what, where and how to evaluate	While we also present the evolution of language models, we focus on the multilingualism of the related task and their evolution alongside the language models

resources or those accommodating different languages.

Delving into the evolution of language models from their preliminary stages to the advent of large language models (LLMs), our survey provides valuable insights from diverse perspectives. While we refrain from delving into details of learning techniques, we aim to provide the broader landscape of multilingual NLP. Table I presents our primary focus and contrasts it with other surveys, outlining our unique contribution to the field. Our survey is mainly designed for people who are knowledgeable about the basics of transfer learning and language models and are interested in applying it to multilingual models and tasks, making it a valuable resource for them.

Structured around the exploration of multilingual models and tasks, the main components of our survey are illustrated in Figure 1. We commence by introducing the fundamental concepts and tracing a brief history of language models, classifying them into main groups in Section II. Subsequently, in Section III, we delve deeper into multilingual models, dissecting their architectures and structures from diverse perspectives. Additionally, we underscore the significance of cross-lingual and multilingual models in NLP, accompanied by insights into available datasets in each application domain, thereby aiding researchers in navigating specific domains within this subject.

Evaluation of these language models often entails analyzing NLP applications that will be present in Section IV, where we review existing literature evaluating models across various languages. Finally, in Section V, we describe future directions and challenges inherent in the subject, offering a comprehensive outlook for future studies.

II. Background

The use of transfer learning in language models has brought about a new phase in Natural Language Processing (NLP). Typically, NLP studies have focused on languages with lots of available data, ignoring those with fewer resources. However, thanks to transfer learning, even languages with limited resources can now be effectively handled. Before we dive into the history of language models and explore transfer learning further, let's first get a basic understanding of some important concepts. In this section, we'll give a simple overview of key ideas like language models and transfer learning.

A. General Concepts

1) Language Models: Language Modeling (LM) stands as a pivotal component in NLP tasks, employing various probabilistic techniques to forecast individual words or sequences within sentences. Its significance in NLP, particularly in the realm of multilingual models, extends to diverse tasks such as machine translation, question answering, speech recognition, and sentiment analysis [10]–[13]. From a statistical perspective, LM entails learning to predict the probability distribution of word sequences in sentences [14], [15]. Through the analysis of text input data, LM acquires insights into the features and characteristics of a language using suitable algorithms, facilitating the understanding of phrases and the prediction of subsequent words in sentences through probabilistic analysis.

2) Transfer Learning: Transfer Learning, a machine learning approach, leverages knowledge gained from pre-training a model on general tasks to enhance efficiency and expedite fine-tuning in other related tasks [8]. This method was first introduced with the advent of ImageNet in 2010, showcasing a successful large Convolutional Neural Network (CNN) model [16]. Through fine-tuning deep neural networks, over 14 million images have been categorized into more than 20,000 classes. Transfer Learning has found extensive application across various NLP tasks and has even yielded state-of-the-art results, particularly in sentiment analysis and other domains.

3) Multilingual and Cross-Lingual Analysis: Multilingual/Cross-lingual learning, often used interchangeably, can be defined as follows:

Multilingual/Cross-lingual learning is a part of transfer learning that focuses on transferring knowledge from one language with usually higher available resources to another

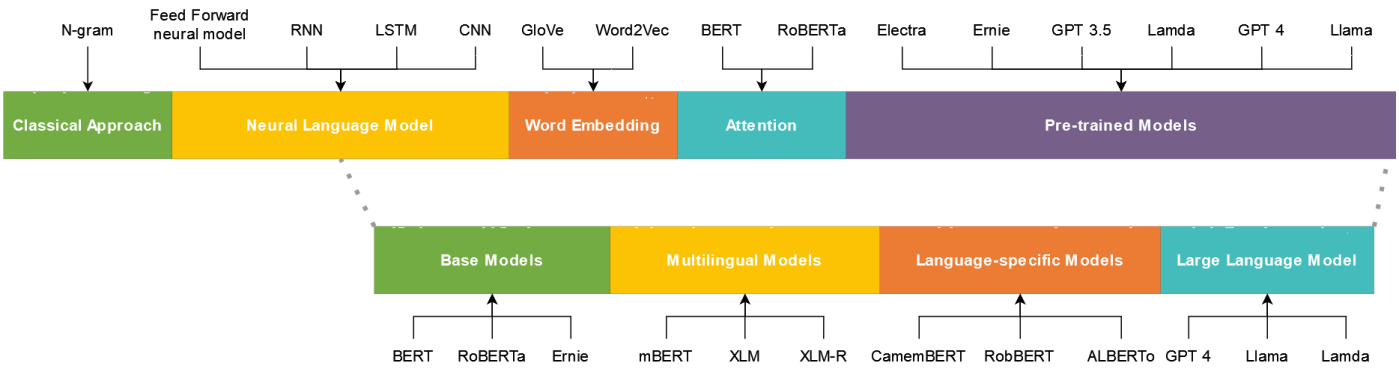

Fig. 2: Evolution of Language Models in NLP

language with lower resources. This concept may lead to better performance in many downstream tasks, especially in languages lacking valuable data. In general, We can look at these concepts from two perspectives:

1: Multilingual usually deals with models. We define this concept as a model pre-trained on different language datasets that check performance on related downstream tasks. Cross-lingual usually comes with learning a model based on a high resource language and then use and evaluate this model for low-resource language for different NLP tasks [5].

2: In terms of cross-lingual embedding, the same vector projection is used for similar words in different languages as a semantic view. In Multilingual embeddings, just using the same embeddings for different languages is considered without assurance of interaction between different languages. In addition, In cross-lingual, we have a query in one language, and the aim is to retrieve the document in another language. However, in Multilingual, in addition to this, the focus is on the models that deal with multiple languages.

4) Zero-shot Learning: Zero-shot Learning (ZSL) involves a classification problem where a classifier is trained on a specific set of labels and then evaluates samples that it hasn't seen before [17].

In multilingual tasks, ZSL refers to classifying data with few or no labeled examples in languages with limited resources, by leveraging training on multiple languages with more available resources. In the context of NLP downstream tasks, ZSL plays a significant role, particularly in cross-lingual applications. For instance, in [18], ZSL is employed for text classification to generalize models on new, unseen classes after training, learning the relationship between sentences and their tag embeddings. Similarly, in news sentiment classification, ZSL is used to assign sentiment categories to news articles in other languages without requiring training data, as demonstrated in [19].

Furthermore, ZSL is applied to question-answering tasks to generalize them to unseen questions, as shown in [20] and [21]. Intent detection, crucial for question-answering, is addressed through zero-shot intent detection, as explored in [22], where user intents are detected for unlabeled utterances. For entity recognition in user talks without annotated data during training, a zero-shot learning approach is presented in [23]. Additionally, [24] analyzes the ZSL approach for Multilingual Sentence Representations in dependency parsing tasks.

B. Neural Language Models

Early methods in NLP research primarily relied on probabilistic language models such as n-grams [25]. These models predict the next word in a sequence by assigning probabilities to word sequences.

In 2001, the first fusion of neural networks with language modeling was proposed [26]. This model improved upon n-gram models by simultaneously learning distributed representations and probability functions for each word, allowing for the use of longer contexts as inputs.

The introduction of Recurrent Neural Networks (RNNs) by Mikolov et al. (2010) marked a significant advancement in NLP. RNNs utilize the output of the previous step as input for predicting the next word, demonstrating remarkable performance in tasks requiring sequential processing. However, due to training challenges, Long Short-Term Memory (LSTM) networks [27] gained popularity for language modeling [28].

Convolutional Neural Networks (CNNs) have also made a notable impact in NLP research. Kalchbrenner et al. (2014) proposed a Dynamic k-Max Pooling network to extract sentence features, offering advantages such as supporting variable-length sentences and applicability to multiple languages. Similarly, Kim (2014) utilized CNNs for sentence-level classification tasks, enhancing performance in tasks like sentiment analysis and question classification.

To address variable-length sequences, Kalchbrenner et al. (2016) introduced ByteNet, incorporating dilation in convolutional layers. Additionally, a combination of CNNs and LSTMs has been used for sentiment analysis [29], while Quasi-Recurrent Neural Networks (QRNNs) were proposed for faster training and testing times compared to LSTMs [30].

In the realm of word embeddings, neural network-based techniques like word2vec [31] and GloVe [32] have gained prominence. Word2vec learns word embeddings using algorithms like Skip Gram and Common Bag of Words (CBOW), while GloVe utilizes unsupervised learning to create embeddings based on word-word co-occurrence probabilities in a large corpus, resulting in improved performance in various

NLP tasks such as named entity recognition and word analogy tasks [32].

C. Pre-trained Language Models

Collobert and Weston introduced a groundbreaking convolutional neural network architecture, serving as a foundational model for pre-trained models in NLP [33]. The output of this architecture for a given sentence can be directly applied to downstream NLP tasks.

The advent of transfer learning heralded a revolution in language model architecture, significantly enhancing performance in downstream NLP tasks. The innovation of bidirectional training in transformers, exemplified by the BERT model [34], enabled training on text sequences in both left-to-right and combined left-to-right and right-to-left directions. In the transformer mechanism, an encoder processes the input text, while a decoder predicts the task's objective. This allows the model to capture context from all preceding and subsequent tokens simultaneously, often resulting in higher accuracy.

The widespread adoption of transfer learning has greatly impacted the development of pre-trained models. It has simplified the process of building NLP models by enabling training on one dataset and then applying the learned knowledge to various NLP tasks on different datasets. This approach is increasingly popular, particularly in multilingual settings, where the structure required for transfer learning aligns well with the demands of multilingualism.

We categorized the existing pre-trained language models into the four main groups:

- *Base Models:* Those types of language models utilized the new architecture and are considered the pioneers of the related structure
- *Multilingual Models:* Those types of language models which deal with multiple languages.
- *Language-specific Models:* Those types of language models which focus on specific languages rather than English
- *Large Language Models:* Those types of language models are trained on massive datasets to process and generate human-like text at scale.

1) Base Models: The term "Base Model" refers to models that garnered significant attention by introducing new structures or altering previous architectures. In our analysis, we primarily focus on BERT and post-BERT models, as illustrated in Figure 2.

In 2018, Google's AI language team introduced a groundbreaking Bidirectional Encoder Representations from Transformers (BERT), revolutionizing the field of pre-trained models. BERT's innovation lies in its ability to jointly learn from unlabeled text in both left and right directions, resulting in remarkable improvements across a wide range of NLP tasks.

A year later, the Facebook AI group introduced a refined method called "RoBERTa" [35], based on BERT's masking strategy but with several key parameter adjustments. Notably, increasing dataset size and training time significantly enhanced performance. RoBERTa also eliminated the "Next Sentence Prediction" task, which was deemed unnecessary.

Another noteworthy model is "ERNIE" (Enhanced Representation through Knowledge Integration), which outperforms Google's BERT in various language tasks, particularly in Chinese [37]. ERNIE focuses on integrating knowledge to enhance representations, leading to improved performance in multilingual contexts.

2) Multilingual Models: With the focus primarily on single language representations, the emergence of multilingual models has garnered significant attention in the field. Following the successful introduction of BERT by Google, a multilingual version was released a year later. Dubbed "mBERT," this model supports sentence representation for 104 languages and has shown superior performance in various multilingual tasks. An analysis of mBERT's semantic aspects by [52] reveals that splitting its representation into language-specific and language-neutral components yields high accuracy, particularly in less challenging tasks such as word alignment and sentence retrieval.

Another notable model based on Transformers and utilizing a masked language modeling (MLM) objective, akin to BERT, is XLM. XLM incorporates translation Language Modeling to learn representations that are similar across different languages [41]. While XLM's structure is rooted in BERT, similar to RoBERTa's parameter adjustments leading to performance improvements, a new multilingual model called XLM-R was introduced. XLM-R removes the translation Language Modeling task and instead employs RoBERTa trained on a larger multilingual dataset encompassing 100 languages [53].

3) Language-Specific Models: While multilingual models have demonstrated high performance across various multilingual tasks, recent research suggests that focusing on a specific language and fine-tuning models for particular tasks in that language can yield even better results in sub-tasks. For instance, the CamemBERT model, a French pre-trained model based on RoBERTa, showcased superior performance by exclusively training on French data and fine-tuning solely for French tasks, outperforming other multilingual models like mBERT and UDify [42].

Table II presents additional language-specific models, underscoring the emerging trend of proposing dedicated models for individual languages in the field of NLP. This approach highlights the importance of tailoring models to specific linguistic contexts to achieve optimal performance.

4) Large language models: Large Language Models (LLMs) represent the latest breakthrough in NLP. These models are predominantly built on deep learning architectures, notably transformer architectures, and are trained on extensive datasets comprising immense amounts of text data. Their advent has brought about significant advancements in NLP, pushing the boundaries of what was previously thought possible. LLMs have facilitated breakthroughs in a myriad of downstream applications including text generation, translation, summarization, and sentiment analysis [54].

III. Multilingual Pre-trained language models

Advancements in transformer efficiency and technology shifts in processing units have paved the way for the de-

TABLE II: Main Characteristics of several existing base, multilingual, language-specific and large language models

Model	Type	Language	Year	Input Corpus Details
BERT [34]	Base model	English	2018	16GB of uncompressed text, BookCorpus (800M words), English Wikipedia (2500M words)
RoBERTa [35]	Base model	English	2019	160GB text: BookCorpus (800M words - 16GB) CC-News (63M English news articles - 76GB), OpenWebText (Web content extracted from URLs shared on Reddit - 38GB), Stories (subset of CommonCrawl data - 31GB)
ELECTRA [36]	Base model	English	2020	For experiments (Same Data as BERT): 3.3 billion tokens from Wikipedia and BooksCorpus. For Language model: extend the BERT dataset to 33B tokens by including data form ClueWeb; CommonCrawl; Gigaword
ERNIE [37]	Base model	English	2020	Processed Wikipedia Eng (4; 500M subwords and 140M entities)
ALBERT [38]	Base model	English	2020	16GB of uncompressed text consists of BookCorpus (800M words) English Wikipedia (2500M words)
UDify [39]	Base model	multilingual	2019	Full universal dependencies v2.3 corpus available on LINDAT, Arabic NYUAD, English ESL, Arabic NYUAD, French FTB, Hindi English HEINCS, Japanese BC-CWJ
XLNet [40]	Base model	English	2019	RACE Dataset, SQuAD, GLUE Dataset, ClueWeb09-B Dataset
mBERT	Multilingual Models	Cross-lingual	2018	Wikipedia, MultiUN, IIT Bombay corpus, OPUS, EUbookshop, OpenSubtitles, GlobalVoices, Kytea and PyThaiNLP5
XLM [41]	Multilingual Models	Cross-lingual	2019	Wikipedia, MultiUN, IIT Bombay corpus, OPUS, EUbookshop, OpenSubtitles, GlobalVoices, Kytea and PyThaiNLP5
CamemBERT [42]	Language-Specific model	French	2019	138GB of uncompressed text and 32.7B SentencePiece tokens consist of: French text extracted from CommonCrawlUnshuffled version of the French OSCAR corpus
RobBERT [43]	Language-Specific model	German	2020	39GB of uncompressed text consists of Dutch Section of OSCAR corpus (6.6B words - 39GB of texts)
BERTje [44]	Language-Specific model	Dutch	2019	Books: a collection of novels (4.4GB), TwNC a Dutch News Corpus (2.4GB), SoNaR-500 reference corpus (2.2GB), 4 Dutch news websites (1.6GB), Wikipedia dump (1.5GB), Total: 12 GB; 2.4B token
ALBERTo [45]	Language-Specific model	Italian	2019	TWITA:from twitter's official streaming API; 200M tweets and 191GB raw data
PhoBERT [46]	Language-Specific model	Vietnamese	2020	20GB texts: Vietnamese Wikipedia corpus (1GB)-(19GB) is a subset of a Vietnamese news corpus
BERT for Finnish [47]	Language-Specific model	Finnish	2019	Yle corpus, an archive of news and STT corpus of newswire articles
ParsBERT [48]	Language-Specific model	Persian	2021	In overall, more that 3M documents from Persian Wikipedia, BigBang Page, Chetor, Eligashtm, Digikala, Ted Talks, books, Miras-Text
GPT-3.5	Large Language model	English	2022	vast amount of text data sourced from various publicly available sources on the internet including websites, books, articles, forums, and other forms of text content across different domains
Lamda [49]	Large Language model	English	2022	comprises 2.97 billion documents, 1.12 billion dialogues, and 13.39 billion dialogue utterances, totaling 1.56 trillion words.
Llama [50]	Large Language model	Multilingual	2023	English CommonCrawl, C4, Github, Wikipedia, Gutenberg and Books3, ArXiv, Stack Exchange
GPT 4 [51]	Large Language model	English	2023	vast amount of text data sourced from various publicly available sources on the internet including websites, books, articles, forums, and other forms of text content across different domains

velopment of language models capable of handling multiple languages simultaneously. In this section, we delve into the significance of multilingual models and assess their level of maturity. We present a comprehensive overview of these models alongside their monolingual counterparts, reviewing their capabilities. Our analysis encompasses research studies from two perspectives: historical evolution and model characteristics. Furthermore, we assess various aspects including architecture, performance metrics, hardware requirements, and language features inherent in these models.

A. Importance of the Multilingual Tasks

Practical applications of NLP often prioritize the English language due to the challenge of training large and accurate language models with small labeled datasets in other languages. However, the importance of developing models for such languages, especially in unforeseen circumstances, has garnered attention. It's worth noting that language models for low-resource languages are not solely limited to emergency situations; they play a crucial role in enabling a wide array of new NLP-dependent technology services. These endeavors

TABLE III: Studies focused on cross-lingual aspects of multilingual models

Title of the Study	Dataset	Evaluation Criteria	Results
How multilingual is Multilingual BERT? [55]	104 languages Wikipedia	examines the multilingual capability	mBERT has an amazing performance in cross-lingual tasks
How Language-Neutral is Multilingual BERT? [52]	use a pre-trained mBERT and train on specific language Wikipedia, WMT14	semantic properties of mBERT	mBERT representations split into a language-specific and a language-neutral component that each one are suitable for specific tasks
Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT [56]	Reuters corpus covering 8 languages	evaluate as a zero-shot cross-lingual model on multiple languages and NLP tasks	fine-tuned hyper parameters mBERT has an amazing performance
Is Multilingual BERT Fluent in Language Generation? [57]	Universal Dependencies treebanks	ability to substitute monolingual models	inefficiency of multilingual models in text generation task
Cross-lingual ability of multilingual BERT: An empirical study [58]	XNLI and LORELEI	cross-lingual ability covering linguistic properties and similarities of languages, model architecture and inputs and training objectives	B-BERT amazing results in cross-lingual applications

are primarily executed within the framework of deep neural networks, highlighting the necessity of language models.

Cross-lingual models leverage large unlabeled datasets in one language to construct a language model, which can then be fine-tuned using a small corpus in another language. This approach significantly enhances performance in the target language, bridging the gap between resource-rich and low-resource languages.

B. Performance Analysis of Multilingual Language Models

In this part, we review the studies that have examined the capabilities of multilingual models. Some focused on the strengths of these models and applications that have good performance; other ones showed NLP tasks in which the performance of multilingual models was inferior to monolingual models. Table III compared these studies.

Pires et al. [55] conducted a comprehensive examination of the multilingual capabilities of the mBERT model. They pre-trained the model on a Wikipedia dataset sourced from over 100 languages, then fine-tuned it with language-specific supervised data for one language, and evaluated its performance on tasks in another language. Their findings revealed that mBERT excels in cross-lingual tasks, with factors such as lexical overlap and typological similarity influencing its performance. Interestingly, the model demonstrated proficiency even in languages with different scripts.

Another study by Libovicky et al. [52] focused on the semantic features of mBERT. They divided the resulting model into two parts: one related to specific languages and the other to general language. While the latter performed well in tasks like word alignment and exact sentence retrieval, it was deemed unsuitable for machine translation applications.

Wu et al. [56] evaluated mBERT as a zero-shot cross-lingual model across approximately 40 languages and five NLP tasks, including natural language inference, document classification, named entity recognition (NER), part-of-speech tagging, and dependency parsing. Their study demonstrated that mBERT achieves excellent performance in these tasks with fine-tuned hyperparameters.

On the contrary, studies such as Ronnqvist et al. [57] have highlighted the inefficiencies of multilingual models in certain applications.

Moreover, research by Karthikeyan et al. [58] delved into BERT's impressive performance in cross-lingual applications, despite lacking a specific cross-lingual objective during training. They investigated the impact of different components of the BERT model on its cross-lingual performance, concluding that factors such as the depth of the network and the total number of parameters in the architecture are more critical than lexical similarity between languages.

Additionally, some studies focus on task-specific optimization of multilingual models, such as the CLBT model by Wang et al. [59], which concentrates on dependency parsing and underscores the influence of lexical properties.

C. Technology Evolution

The process of technology development in the field of multilingual models can be studied from a historical perspective (evolution in time) or a model perspective, which will be detailed in this section.

1) Historical Review: In terms of historical evolution over time, the development of multilingual models has traversed a challenging trajectory (Figure 3). Initially, models like ELMO [60] adhered to bidirectional LSTM architectures and exhibited commendable performance. However, the introduction of transformers [61] marked a significant shift, dominating the architecture and performance landscape of multilingual models for a period. Transformers revolutionized model architecture by replacing recursive structures with attention mechanisms, thereby enhancing parallel execution and performance across various tasks. Nevertheless, this transition also escalated the demand for processing resources and extended training times.

Subsequently, the release of the BERT model heralded a new era, prompting further refinements and advancements

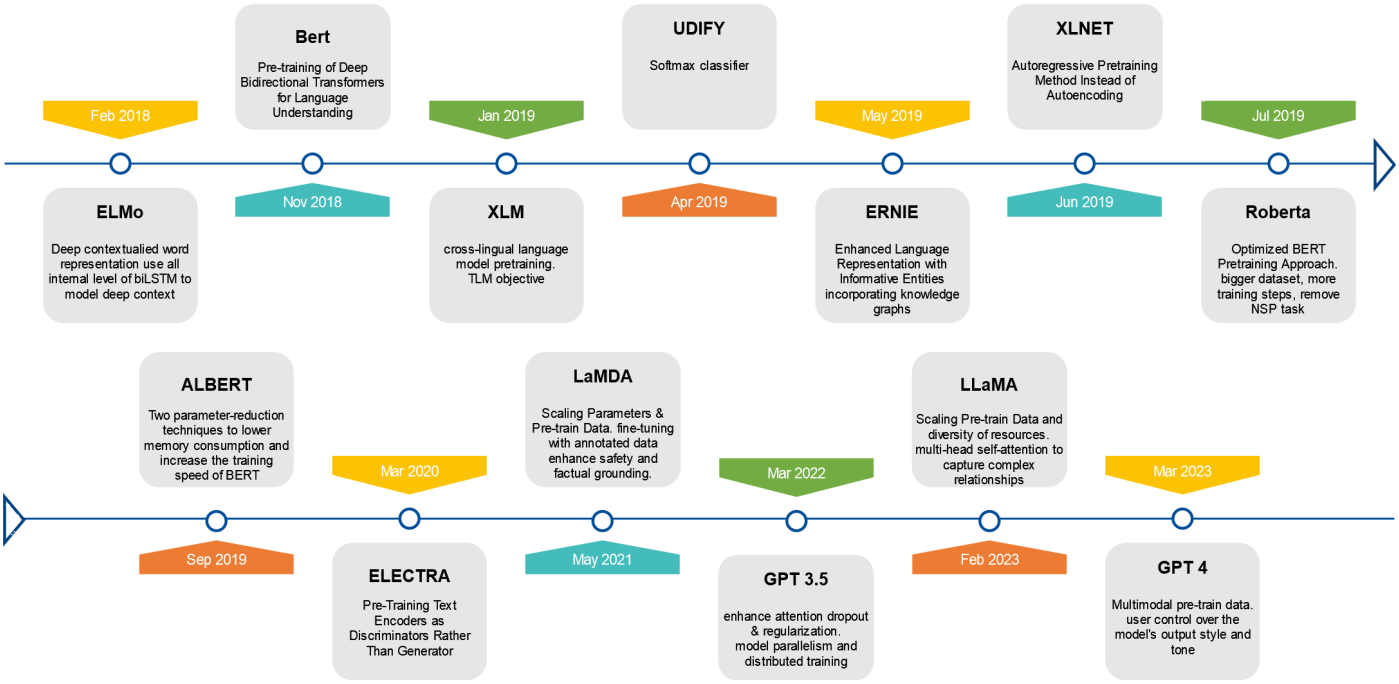

Fig. 3: Evolution of Linguistic Technologies (from Time Perspective)

in subsequent works. For instance, the ALBERT model [38] employed techniques to reduce parameter counts while maintaining the performance of large BERT models, resulting in more lightweight versions.

As time progressed, researchers pursued divergent paths in design and architecture, introducing novel models such as XLNet [40], which leveraged an autoregressive model, and ELECTRA [36], which innovatively pre-trained a text encoder as the generator. These innovative approaches have continued to push the boundaries of multilingual model development.

The most recent trend in the field, from a time prespective, revolves around LLMs. This emergence is marked by the introduction of groundbreaking models like LaMDA [49], Llama [50], and GPTs [51].

2) Model Perspective: From the model point of view, according to Figure 4, we divided our studies into four categories:

The first generation that came before introducing BERT, such as ELMO [60], shifted the results by using all the output of the Bidirectional LSTM inner layers.

In the second generation, BERT and its minor improvements are categorized, using more extensive data sets and changing pre-train tasks and classifier optimizations are major changes seen in models such as:

mBERT: Multilingual BERT published same time as BERT, supports over 100 languages. Technically, It is just BERT trained on Wikipedia text of many languages. For the content size bias resistance for different languages, low-resource languages were oversampled and general languages were undersampled.

UDify: This model uses over 120 Universal Dependencies [62] treebanks in more than 70 languages and fine-tuned BERT on all datasets as a single one. That shows state-of-the-art universal POS, UFeats, Lemmas, UAS, and LAS scores. Hence can be assumed, multilingual multi-task model. [39]

XLM: This study was presented to evaluate Pre-trained cross-lingual models (XLMs) and suggested two methods for pre-training. The first method is unsupervised pre-training based on monolingual data, and the second method is pre-training based on multilingual data. Evaluations were performed in the XLNI [63] and WMT'16 tasks [64]. Another innovation of this research [41] is the introduction of several objectives for pre-learning. They used MLM and Causal Language Modeling (CLM) for unsupervised learning, which examined its proper performance. They also used translation language modeling objective (TLM) alongside MLM, which is essentially an extension of MLM in the BERT model, using a set of parallel sentences instead of consecutive sentences.

XLM-R: A self-supervised model uses RoBERTa objective task on a CommonCrawl dataset[1] contains the unlabeled text of 100 languages with a token number of five times more than RoBERTa. The advantage of this model is that, unlike XLM, it does not require parallel entry, so it is scalable. [53]

In the Post-BERT era, models had significant modifications, for instance, using an auto-regressive pre-train instead of an auto-encoder. As an example, XLNet, focuses on autoregressive models that attempt to estimate the probability distribution of the text. In contrast, autoencoding models such as BERT try to reconstruct the original data by seeing incomplete data generated by covering some sentence tokens. Other models took a relatively different path than the BERT-based models. For example, such as ELECTRA, that the encoder is trained as a discriminator instead of a generator. However,

[1] https://commoncrawl.org

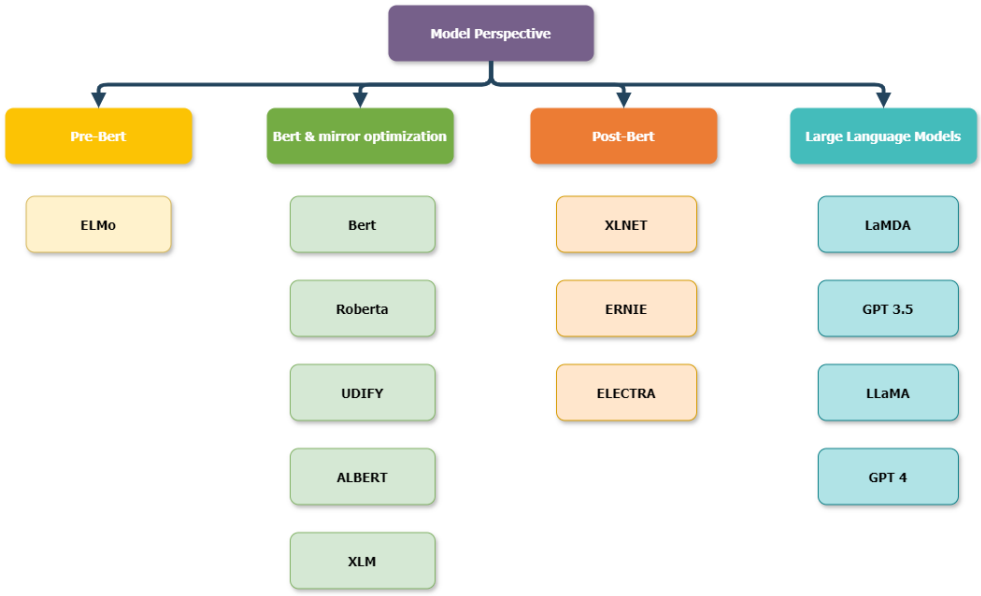

Fig. 4: Evolution of Linguistic Technologies (from Modeling Perspective)

GPT [65] and mBERT [34] focus on learning contextual word embeddings. These learned encoders are still needed to represent words in context by downstream tasks. Besides, various pre-training tasks are also proposed to learn PTMs for different purposes.

Finally, with the emergence of the term "Large Language Models," notable examples include LaMDA [49], a family of conversational LLMs developed by Google; LLama [50], an autoregressive LLM released by Meta AI; and GPT [51] by OpenAI. These models are characterized by their substantial increase in parameters and input corpora, marking a significant advancement in the field.

D. Architecture and Hardware Requirements

From an architectural standpoint, the majority of models follow a structure similar to BERT-base or BERT-large [36] [38] [40]. These models typically employ a combination of transformer and attention layers, with attention layers playing a crucial role in capturing the meaning and context of words.

However, there are exceptions where models deviate from the standard BERT architecture [37] or adopt alternative approaches [60] [41].

Furthermore, batch size serves as another point of comparison among models. Models akin to BERT often utilize a bigger batch size while others like ERNIE opt for a smaller size of 512. This variation in batch size can impact training efficiency and resource utilization.

In this section, we evaluate the efficiency of various models in terms of hardware requirements.

As depicted in Table IV, different research teams have introduced foundational models based on the Transformers architecture. These models vary in terms of their architecture, total number of model parameters, and the hardware platforms they utilize.

The hardware processing units employed by these models predominantly include TPUs (Tensor Processing Units) or GPUs (Graphics Processing Units). While each model may utilize a proprietary combination of hardware resources, some instances stand out, such as XLNet, which employed up to 512 TPUs for less than three days. Notably, according to the CEO of Hologram AI, this endeavor incurred a cost of $245,000 and produced 5 tons of CO2 emissions. Such substantial investments were made to surpass BERT in 18 out of 20 tasks [68].

In terms of the number of model parameters, there is considerable variation among models. For instance, the ELECTRA model's smallest version contains 14 million parameters, whereas the ALBERT Large model boasts 235 million parameters. This diversity in parameter count reflects the range of complexities and capabilities exhibited by these models.

E. Datasets

In table V, several available datasets using various languages are introduced, and for each dataset, in addition to a short description, we provided the evaluation metrics and the task that was used in previous studies.

F. Multilingual Tasks Domains

As shown in Figure 5, we can categorized linguistic domain to be considered for multilingual tasks from several perspectives:

- *From Morphology point of view*: Since morphology deals with the formation of words and the relation of words together, defining this category is meaningful in the multilingual task because this formation varies in different languages but can have many common properties too. The morphological structure of words usually consists of prefixes/suffixes, singularization/pluralization, gender detection, word inflection (words modification in order to express grammatical categories).

TABLE IV: Architecture and Hardware requirements of several models

Model	Team	Architecture Details	Params Number	Hardware
BERT [34]	Google AI	Based on the Transformer architecture; deeply bidirectional model Base:12 layers (transformer blocks); 12 attention headsLarge: 24 layers (transformer blocks); 16 attention heads	Base:110M Large:335M	4 to 16 Cloud TPUs; 1 TPU; 64 gb ram
ELECTRA [36]	Stanford University; Google Brain	Transformers (Same as BERT) -Generators and Discriminators- ELECTRA-small: 256 hidden dimensions (instead of 768); 128 token embedding (instead of 768); 128 sequence length (instead of 512)	Small:14M Large:110M	Small : 1 V100 GPU Large: 16TPUv3s
ERNIE [37]	Tsinghua University, Huawei Noah's Ark Lab	BERT + two multi-head self-attention. 6 layer textual encoder, 6 layer knowledgeable encoder, hidden dimension of token embedding=768, hidden dimension of entity embedding= 100, self-attention heads: Aw = 12, Ae = 4	114M	8 NVIDIA-2080Ti
ALBERT [38]	Google Research; Toyota Technological Institute at Chicago	4 models: base with 12 layers and 768 hiddens, large with 24 layers and 1024 hiddens, xlarge with 24 layers and 2048, xxlarge with 24 layers and 4096 hiddens	Base:12M Large:18M XL:60M XXL:235M	64 to 512 Cloud TPU V3
ELMo [60]	Allen Institute; Allen School of CS; University of Washington	2 BiLSTM layers with 4096 units and 512 dimension projections and a residual connection from the first to second layer	499M [66]	3 GTX 1080 [67]
XLM [41]	Facebook AI Research	1024 hidden units, 8 heads, GELU activation	XLM-15:250M XLM-17:570M XLM-100:570M	64 Volta GPUs for the language modeling tasks, and 8 GPUs for the MT tasks
XLNet [40]	Carnegie Mellon University; Google AI Brain Team	same as BERT-Large, batch size of 8192	110M	512 TPU v3

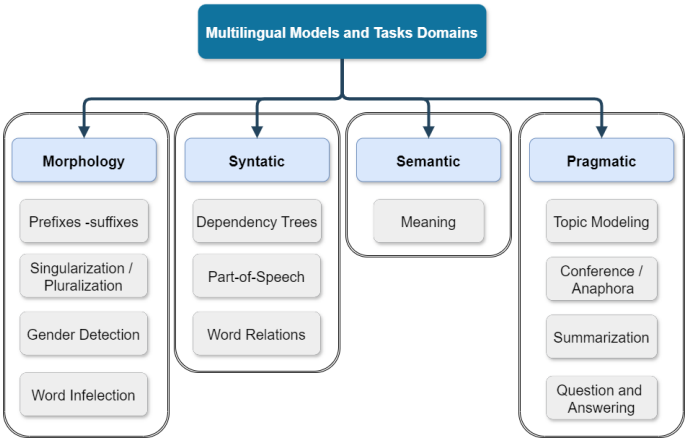

Fig. 5: Multilingual Tasks Domains

- *From Syntax point of view*: Syntactic perspective in multilingual tasks refers to words relation and combination to form a bigger language unit such as sentences, clauses, and phrases. In everyday life, this view is more commonly known as the grammatical view. Alongside the relation between words, part of speech and dependency tree are considered in this category.

- *From Semantics point of view*: This view refers to the meaning of words and sentences. Semantic perspective is one of the main categories in linguistics for the multilingual task because semantic structure and relation of words and sentences are essential features of any language.

- *From Pragmatics point of view*: The pragmatic perspective in multilingual tasks deals with the contribution of context to meaning. Several hot topics such as topic modeling, coreference-anaphora, summarization, and question-answering in NLP are considered in this perspective.

IV. Multilingual Applications and Tasks

With the vast amount of data generated daily across various forms, including unstructured data, emails, chats, and tweets, the significance of NLP tasks and applications continues to grow. Leveraging these applications for data analysis enables businesses to derive valuable insights. Notably, trending topics such as elections and Covid-19 often drive heightened content generation on social media platforms, necessitating attention from the NLP community. While analyzing data for low-resource languages poses challenges, NLP has proven successful across various applications, including virtual assistants, speech recognition, sentiment analysis, and chatbots [103]. For instance, Google Translate, a free multilingual machine translation service developed by Google, relies on NLP in its operations. Similarly, Amazon Alexa and Google Assistant employ speech recognition and NLP techniques, such as question answering, text classification, and machine translation, to assist users in achieving their objectives. Even within the digital marketing sector, utilizing these techniques for data analysis aids in understanding customer interests and generating precise reports tailored to business requirements.

The primary aim of this research is to comprehensively investigate the various tasks and applications within NLP, extending beyond dominant languages such as English. To facilitate the progression of research and considering the abundance of NLP models and applications aimed at supporting multiple languages, it is imperative to identify and review the existing research conducted in this domain.

TABLE V: Details of available dataset for various tasks

Dataset Name	Description	Task	Languages	Used in
SNIPS [69]	contains several day to day user command categories (e.g. play a song, book a restaurant).	Slot Filling and Intent Detection	EN DE ES FR IT JA KO PT_BR PT_PT	[70] [71] [72] [73]
MTOP [74]	parallel multilingual task-oriented semantic parsing corpora. crowd-sourced 100k examples in 11 domains and 117 intents used for 3-way evaluation: in-language, multilingual, zero-shot	Slot Filling and Intent Detection	EN DE FR ES HI TH	[75] [76]
Multilingual ATIS [77]	ATIS dataset subset translated to 2 languages by a human expert to show that results can surpass the proposed approaches with only a few labeled tokens.	Slot Filling and Intent Detection	EN HI TR	-
Facebook's multilingual db [78]	under 60k annotated utterances about alarm-reminder and weather	Slot Filling and Intent Detection	EN TH ES	[79]
CommonCrawl	Over petabyte crawled web data from 2008 and released it publicly	MLM	more than 40 languages	[53] [80] [81] [82]
MultiATIS++ [83]	extend train and test set of the English ATIS	Slot Filling and Intent Detection	EN ES DE FR PT HI ZH JA TR	-
XED [84]	A multilingual fine-grained emotion dataset	Sentiment analysis	Mainly EN , Finnish and 30 additional languages	-
WikiAnn [85]	cross-lingual name tagging and linking based on Wikipedia articles. Assigning a coarse-grained or fine-grained type to each mention, and link it to an English Knowledge Base if it is linkable	NER	295 languages	[86] [87]
CODAH [88]	an evaluation dataset with 2.8k questions for testing common sense. Model challenging extension to the SWAG dataset, which tests commonsense knowledge using sentence-completion questions that describe situations observed in video.	Question Answering	EN	[89] [90] [91]
HotpotQA [92]	dataset with 113k Wikipedia-based question-answer pairs	Question Answering	EN	[93] [94] [95]
NewsQA [96]	reading comprehension dataset of over 100K human-generated question-answer pairs from over 10K news articles from CNN, with answers consisting of spans of text from the corresponding articles	Question Answering	EN	[95] [97] [93]
GoEmotions [98]	dataset of 58k English Reddit comments, labeled for 27 emotion categories or Neutral	Sentiment Analysis, Emotion Analysis	EN	[99] [100] [101] [102]

Transfer learning emerges as a valuable tool in achieving high performance across numerous NLP tasks, not only in well-resourced languages like English but also in many low-resource languages. As non-English language models gain traction in both academic and industrial spheres, recent research endeavors increasingly emphasize the multilingual aspect of NLP across various tasks. Moreover, in certain scenarios, there exists more of a cross-domain advantage than a strictly multilingual advantage. Transfer learning from a pre-trained multilingual model to a language-specific model can significantly enhance performance across various downstream tasks. Kuratov et al. [104] exemplify this approach with the Russian language, showcasing performance improvements in reading comprehension, paraphrase detection, and sentiment analysis tasks, alongside a reduction in training time compared to multilingual models.

An essential aspect of text analysis involves examining the style of the text. Numerous factors, including formality markers, emotions, and metaphors, play pivotal roles in influencing the analysis of textual style. Kang et al. [105] contribute to this domain by providing a benchmark corpus (xSLUE) comprising text in 15 different styles and 23 classification tasks, serving as an online platform for cross-style language understanding and evaluation. This research underscores the diverse avenues available for developing low-resource or low-performance styles and other applications, such as cross-style generation .

Another significant challenge in NLP applications, particularly in low-resource languages, pertains to the detection of hate speech [106].

Also an architecture for pre-trained transformers aimed at exploring cross-lingual zero-shot and few-shot learning [106]. Their model incorporates the innovative attention-based classification block AXEL, leveraging transformer techniques on both English and Spanish datasets.

Moreover, Tawalbeh et al. [107] utilized transfer learning with BERT and RNN to address shared tasks concerning multilingual offensive language detection.

A. Translation

The significance of translation in the realm of NLP is indisputable, particularly concerning multilingual contexts where this service takes center stage. Many of these models are trained primarily on a single language, typically English, and endeavor to translate into other languages. Notably, Facebook AI introduced "M2M-100" [108], a Many-to-Many multilingual translation model capable of translating directly between any pair of languages from a pool of 100 languages.

Employing zero-shot systems, authors in [109] delve into the proximity between languages, focusing on both automatic standard metrics such as BLEU and TER .

B. Speech Recognition

Much research has been conducted in the field of Speech Recognition, primarily emphasizing deep neural networks and Recurrent Neural Networks (RNNs) [110]–[112]. However, with the burgeoning adoption of transformers in NLP, recent research endeavors in the domain of speech recognition predominantly integrate transformers into their architectures. For multilingual speech recognition, Zhou et al. [113] introduced a sequence-to-sequence attention-based model featuring a single Transformer that employs sub-words without relying on any pronunciation lexicon for their model.

C. Sentiment Analysis

Sentiment analysis aims to discern and extract information such as feelings, attitudes, emotions, and opinions from textual content. Many businesses leverage this service to enhance their product quality by scrutinizing customer feedback. However, a primary challenge lies in achieving satisfactory performance for languages with limited resources. To address this, Can et al. [114] trained a model on a high-resource language (English) and repurposed it for sentiment analysis in other languages (Russian, Spanish, Turkish, and Dutch) with less abundant data while in [115], they proposed a language-agnostic method for sentiment classification and evaluated by approaches based on four deep models. Authors in [116] proposed a novel deep learning method addressing the significant challenges in multilingual sentiment analysis, aiming to mitigate excessive reliance on external resources. Additionally, Kanclerz et al. [117] introduced a novel technique utilizing language-agnostic sentence representations to adapt a model trained on texts in Polish (a low-resource language) for recognizing polarity in texts in other languages with higher resource availability . These efforts signify strides toward overcoming the challenges inherent in multilingual sentiment analysis.

D. Intent detection and Slot filling

Intent Detection involves identifying the user's current goal and assigning appropriate labels, commonly employed in chatbots and intelligent systems. Conversely, slot filling aims to extract attribute values of specific types. Studies indicate a strong correlation between these two tasks, often resulting in achieving state-of-the-art performance [118], [119]. Models in this domain typically leverage joint deep learning architectures within attention-based recurrent frameworks. Castellucci et al. [120] and researchers in [79] proposed a "recurrence-less" model utilizing BERT-Join, which demonstrated robust performance for these tasks. Notably, they achieved similar performance for the Italian language without necessitating model adjustments.

E. Dependency Parsing

Dependency parsing poses a significant challenge, particularly in multilingual NLP. Wang et al. [59] tackled this challenge by employing the BERT transformation approach to generate cross-lingual contextualized word embeddings. Through a linear transformation learned from contextual word alignments trained across various languages, their method demonstrated effectiveness in zero-shot cross-lingual transfer parsing. Furthermore, their approach showcased superiority over static embeddings.

F. NER

Named Entity Recognition (NER) involves extracting entities from text and categorizing them into predefined categories. Recent advancements in self-attention models have demonstrated state-of-the-art performance in this task, particularly for inputs comprising multiple sentences. This capability becomes increasingly vital when analyzing data across multiple languages. Luoma et al. [121] leveraged BERT in five languages to explore the utilization of cross-sentence information for NER, showcasing superior performance across all tested languages and models.

In scenarios where languages possess limited or no labeled data, Wu et al. [122] proposed a teacher-student learning method to address this challenge in both single-source and multi-source cross-lingual NER.

Moreover, for assessing different architectures in the task of name transliteration within a many-to-one multilingual paradigm, including LSTM, biLSTM, GRU, and Transformer, Moran et al. [123] demonstrated enhanced accuracy with the transformer architecture for both encoder and decoder components.

G. Question Answering

Question Answering (QA) involves developing an automated system to respond to questions posed by humans in natural language [124]. This task is receiving considerable attention, particularly in the realm of multilingualism, yet it remains highly challenging. Different languages employ diverse approaches to constructing meaning. For instance, in English, the plural form of words often involves adding an 's' at the end, whereas in Arabic, forming plurals may entail more complex structural changes rather than simply adding postfixes to words. Additionally, languages like Japanese may not utilize spaces between words [125]. These linguistic intricacies underscore the complexity of multilingual QA systems.

V. Challenges and Outlook

This section provides some of the challenges in the domain of multi-lingual tasks and a set of ideas to be considered as future direction of this research line.

A. Existing Challenges

We identified three groups of challenges in the domain of using transfer learning for multilingual tasks including challenges on (i) Modeling, (ii) practical aspects and (iii) applications. Next, we provide details on each group of challenges.

1) Modeling: challenges of pre-trained models due to the complexity of natural language processing can be grouped as follows:

- Various objective tasks that evaluate different features of models. A challenging objective task can help in the manner of creating more general models. However, these tasks should be self-supervised because many captured corpora do not have tagged data.
- Due to the increasing use and research on multilingual and cross-lingual models, their vulnerability and reliability have become very important. In Section III-B, we reviewed some researches in this area and noted the less studied multilingual models. Nowadays, most of the researches in this category, conducted on mBERT.

2) Practical: Research studies on following problems are affected by the high cost of pre-training models:

- General purpose models can learn the fundamental understanding of languages. However, usually need more profound architecture, larger corpora, and Innovative pre-training tasks.
- Recent studies have confirmed the performance of Transformers in pre-trained models. Nevertheless, the computational resource requirement of these models limits their application. Therefore, model architecture improvement needs more attention in the research area. Moreover, architecture improvements could lead to a better contextual understanding of the language model, as it could deal with a more extended sequence and recognize context. [126]
- Achieve maximum performance of current models: Most existing models can improve performance with increasing model depth, for example, with a more comprehensive input corpus or train steps.

3) Application:

- In terms of multilingual tasks, many task do not have enough data resources to gain significant performance in a specific application.
- The next big challenge is to successfully execute NER, which is essential when training a machine to distinguish between simple vocabulary and named entities. In many instances, these entities are surrounded by dollar amounts, places, locations, numbers, time, etc., it is critical to make and express the connections between each of these elements, only then may a machine fully interpret a given text.
- Another challenge to mention is extracting semantic meanings. Linguistic analysis of vocabulary terms might not be enough for a machine to correctly apply learned knowledge. To successfully apply learning, a machine must understand further, the semantics of every vocabulary term within the context of the document.

B. Future Directions

This study offers insights into the future directions of research within the multi-lingual tasks domain. The following avenues can be considered for further exploration:

- Vertical Extension: Enhancing the performance of current models through increased pre-training steps, parameters, and input corpora size. However, this necessitates higher processing power, highlighting the need to analyze the relationship between hyperparameters and model performance.
- Horizontal Expansion: Expanding research studies with multilingual corpora pre-training and evaluation across various downstream tasks can lead to improved model performance. Similar to vertical extensions, this requires substantial processing resources.
- Pre-training Tasks: Investigating pre-training tasks, particularly in cross-lingual models, presents a challenging yet promising research field. Advancements in this area can lead to more comprehensive model evaluations.
- Optimization of Model Architecture: Deepening research into model architecture design and training methods can yield models capable of pre-training on vast multilingual corpora with existing computing resources.
- Specialized Purpose Models: There is a growing trend towards developing models tailored for specific domains such as health advice. However, there remains a gap in addressing low-resource or real-time computing needs. Designing models with specific pre-training objectives for such tasks is essential.
- Robustness: Ensuring the robustness of pre-trained models requires further attention. Studies focusing on this aspect will offer valuable insights into the future deployment of these models in various industries.
- Recent advancement in LLMs: Further investigation into recent advances in large language models is essential. By closely examining these advancements, researchers can glean valuable insights into pushing the boundaries of multi-lingual tasks. Integrating the latest findings from the realm of large language models into ongoing research efforts will undoubtedly enrich the understanding and capabilities of future models.

By addressing these areas, researchers can advance the field of multi-lingual tasks and contribute to the development of more efficient and effective language models.

VI. Conclusion

This survey offers a comprehensive overview of existing studies of the evolution of language models to address multilingual and cross-lingual tasks. In addition to reviewing various models, we also examined the primary datasets available in

the community and explored different approaches in terms of architectures and applications. Through this analysis, we identified several research challenges within the domain. Subsequently, we propose several potential future directions to advance research in this field.

VII. Acknowledgments

A sincere thanks to Prof. Mahdi Jalili for his helpful insights for this paper.

References

[1] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[2] S. Ruder, I. Vulić, and A. Søgaard, "A survey of cross-lingual word embedding models," *Journal of Artificial Intelligence Research*, vol. 65, pp. 569–631, 2019.

[3] A. Malte and P. Ratadiya, "Evolution of transfer learning in natural language processing," *arXiv preprint arXiv:1910.07370*, 2019.

[4] R. Dabre, C. Chu, and A. Kunchukuttan, "A survey of multilingual neural machine translation," *arXiv*, vol. 53, no. 5, 2019.

[5] M. Pikuliak, M. Šimko, and M. Bieliková, "Cross-lingual learning for text processing: A survey," *Expert Systems with Applications*, vol. 165, 2021.

[6] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, 2023.

[7] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.

[8] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 15–18.

[9] M. Pikuliak, M. Šimko, and M. Bieliková, "Cross-Lingual Learning for Text Processing: A Survey," *Expert Systems with Applications*, vol. 165, p. 113765, aug 2020.

[10] A. Vaswani, Y. Zhao, V. Fossum, and D. Chiang, "Decoding with large-scale neural language models improves translation," *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1387–1392, 2013.

[11] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, "Question Answering Systems: Survey and Trends," *Procedia Computer Science*, vol. 73, no. Awict, pp. 366–375, 2015.

[12] T. Mikolov, M. Karafiát, L. Burget, C. Jan, and S. Khudanpur, "Recurrent neural network based language model," *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, no. September, pp. 1045–1048, 2010.

[13] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27, no. February, pp. 16–32, 2018.

[14] S. Osborne, "Learning NLP Language Models with Real Data," 2019, Accessed: 2021-06-28. [Online]. Available: https://towardsdatascience.com/learning-nlp-language-models-with-real-data-cdff04c51c25

[15] N. A. Smith, "Probabilistic Language Models 1.0," Tech. Rep., 2017.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[17] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks," *Proceedings of the National Conference on Artificial Intelligence*, vol. 2, pp. 646–651, 2008.

[18] P. K. Pushp and M. M. Srivastava, "Train once, test anywhere: Zero-shot learning for text classification," *CoRR*, vol. abs/1712.05972, 2017.

[19] A. Pelicon, M. Pranjić, D. Miljković, B. Škrlj, and S. Pollak, "Zero-shot learning for cross-lingual news sentiment classification," *Applied Sciences (Switzerland)*, vol. 10, no. 17, 2020.

[20] K. Ma, F. Ilievski, J. Francis, Y. Bisk, E. Nyberg, and A. Oltramari, "Knowledge-driven Self-supervision for Zero-shot Commonsense Question Answering," *CoRR*, no. Lm, 2020.

[21] P. Banerjee and C. Baral, "Self-supervised Knowledge Triplet Learning for Zero-shot Question Answering," *arXiv*, pp. 151–162, 2020.

[22] C. Xia, C. Zhang, X. Yan, Y. Chang, and P. S. Yu, "Zero-shot user intent detection via capsule neural networks," *arXiv*, pp. 3090–3099, 2018.

[23] M. Guerini, S. Magnolini, V. Balaraman, and B. Magnini, "Toward zero-shot entity recognition in task-oriented conversational agents," *SIGDIAL 2018 - 19th Annual Meeting of the Special Interest Group on Discourse and Dialogue - Proceedings of the Conference*, no. July, pp. 317–326, 2018.

[24] K. Tran and A. Bisazza, "Zero-shot dependency parsing with pre-trained multilingual sentence representations," *arXiv*, pp. 281–288, 2019.

[25] R. E. Banchs, J. M. Crego, P. Lambert, and M. R. Costa-juss, "N -gram-based Machine Translation," in *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, no. April, 2006.

[26] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in Neural Information Processing Systems*, 2001.

[27] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] A. Graves, "Generating Sequences With Recurrent Neural Networks," pp. 1–43, 2013.

[29] J. Wang, L. C. Yu, K. R. Lai, and X. Zhang, "Dimensional sentiment analysis using a regional CNN-LSTM model," *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Short Papers*, pp. 225–230, 2016.

[30] J. Bradbury, S. Merity, C. Xiong, and R. Socher, "Quasi-recurrent neural networks," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–11, 2017.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, pp. 1–12, 2013.

[32] C. D. M. Jeffrey Pennington, Richard Socher, "GloVe: Global Vectors for Word Representation," *British Journal of Neurosurgery*, vol. 31, no. 6, pp. 682–687, 2017.

[33] R. Collobert and J. Weston, "A unified architecture for natural language processing," pp. 160–167, 2008.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," no. Mlm, 2018.

[35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," no. 1, 2019.

[36] C. D. Manning, "Electra : PRe-Training Text Encoders As Discriminators Rather Than Generators," *Iclr*, pp. 1–18, 2020.

[37] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE : Enhanced Language Representation with Informative Entities," 2019.

[38] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite Bert for self-supervised learning of language representations," pp. 1–17, 2020.

[39] D. Kondratyuk and M. Straka, "75 Languages, 1 Model: Parsing Universal Dependencies Universally," pp. 2779–2795, 2019.

[40] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," no. NeurIPS, pp. 1–18, 2019.

[41] G. Lample and A. Conneau, "Cross-lingual Language Model Pretraining," 2019.

[42] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a Tasty French Language Model," vol. 2, 2019.

[43] P. Delobelle, T. Winters, and B. Berendt, "RobBERT: a Dutch RoBERTa-based Language Model," 2020.

[44] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim, "BERTje: A Dutch BERT Model," 2019.

[45] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile, "AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets," *CEUR Workshop Proceedings*, vol. 2481, 2019.

[46] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," no. February, pp. 14–16, 2020.

[47] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, and S. Pyysalo, "Multilingual is not enough: BERT for Finnish," 2019.

[48] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, "Parsbert: Transformer-based model for persian language understanding," *Neural Processing Letters*, vol. 53, no. 6, pp. 3831–3847, 2021.

[49] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du *et al.*, "Lamda: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022.

[50] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[51] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[52] J. Libovický, R. Rosa, and A. Fraser, "How Language-Neutral is Multilingual BERT?" 2019.

[53] Alexis Conneau, Kartikay Khandelwal, Naman Goyal Vishrav, and Vishrav Chaudhary, "Unsupervised Cross-Lingual Representation Learning at Scale," pp. 31–38, 2019.

[54] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[55] T. Pires, E. Schlinger, and D. Garrette, "How Multilingual is Multilingual BERT?" pp. 4996–5001, 2019.

[56] S. Wu and M. Dredze, "Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 833–844.

[57] S. Rönnqvist, J. Kanerva, T. Salakoski, and F. Ginter, "Is Multilingual BERT Fluent in Language Generation?" 2019.

[58] K. Karthikeyan, Z. Wang, S. Mayhew, and D. Roth, "Cross-lingual ability of multilingual bert: An empirical study," dec 2019.

[59] Y. Wang, W. Che, J. Guo, Y. Liu, and T. Liu, "Cross-lingual BERT transformation for zero-shot dependency parsing," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 5721–5727, 2020.

[60] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep Contextualized Word Representations," 2018, pp. 2227–2237.

[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," pp. 5998–6008, 2017.

[62] J. Nivre, M. Abrams, Ž. Agić, and Ahrenberg, "Universal dependencies 2.3," 2018, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. [Online]. Available: http://hdl.handle.net/11234/1-2895

[63] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, "XNLI: evaluating cross-lingual sentence representations," *CoRR*, vol. abs/1809.05053, 2018.

[64] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318.

[65] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," *OpenAI*, pp. 1–10, 2018.

[66] L. H. Li, P. H. Chen, C. J. Hsieh, and K. W. Chang, "Efficient contextual representation learning without softmax layer," feb 2019.

[67] matt peters, "no.of GPUs used for training 1 Billion Word Benchmark?" 2018. [Online]. Available: https://github.com/allenai/bilm-tf/issues/55

[68] Synced, "The Staggering Cost of Training SOTA AI Models," 2019. [Online]. Available: https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

[69] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril *et al.*, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, pp. 12–16, 2018.

[70] A. Babu, A. Shrivastava, A. Aghajanyan, A. Aly, A. F. Marjan, and G. Facebook, "Non-Autoregressive Semantic Parsing for Compositional Task-Oriented Dialog," Tech. Rep.

[71] Q. Chen, Z. Zhuo, and W. Wang, "BERT for Joint Intent Classification and Slot Filling," Tech. Rep.

[72] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, M. Primet, and J. Dureau, "Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces," Tech. Rep.

[73] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. Jeff Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB: Speech processing Universal PERformance Benchmark," Tech. Rep., 2021.

[74] H. Li, A. Arora, S. Chen, A. Gupta, S. Gupta, and Y. Mehdad, "MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark," 2020.

[75] S. Desai, A. Shrivastava, A. Zotov, and A. Aly, "Low-Resource Task-Oriented Semantic Parsing via Intrinsic Modeling," apr 2021.

[76] P. Kaliamoorthi, A. Siddhant, E. Li, and M. Johnson, "Distilling Large Language Models into Tiny and Effective Students using pQRNN," 2021.

[77] S. Upadhyay, M. Faruqui, G. Tür, H.-T. Dilek, and L. Heck, "(almost) zero-shot cross-lingual spoken language understanding," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6034–6038.

[78] S. Schuster, R. Shah, S. Gupta, and M. Lewis, "Cross-lingual transfer learning for multilingual task oriented dialog," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 3795–3805, 2019.

[79] Z. Zhang, Z. Zhang, H. Chen, and Z. Zhang, "A joint learning framework with bert for spoken language understanding," *IEEE Access*, vol. 7, pp. 168 849–168 858, 2019.

[80] B. Myagmar, J. Li, and S. Kimura, "Cross-domain sentiment classification with bidirectional contextualized transformer language models," *IEEE Access*, vol. 7, pp. 163 219–163 230, 2019.

[81] B. Tahir and M. A. Mehmood, "Corpulyzer: A novel framework for building low resource language corpora," *IEEE Access*, vol. 9, pp. 8546–8563, 2021.

[82] Z. Li, X. Li, J. Sheng, and W. Slamu, "Agglutifit: Efficient low-resource agglutinative language model fine-tuning," *IEEE Access*, vol. 8, pp. 148 489–148 499, 2020.

[83] W. Xu, B. Haider, and S. Mansour, "End-to-End Slot Alignment and Recognition for Cross-Lingual NLU," pp. 5052–5063, 2020.

[84] K. Kajava and J. Tiedemann, "XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection," Tech. Rep.

[85] X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, and H. Ji, "Cross-lingual name tagging and linking for 282 languages," in *Proceedings*

of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1946–1958.

[86] E. M. Ponti, I. Vuli´cvuli´c, R. Cotterell, P. Parovi´c, R. Reichart, and A. Korhonen, "Parameter Space Factorization for Zero-Shot Learning across Tasks and Languages," Tech. Rep.

[87] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel, "ByT5: Towards a token-free future with pre-trained byte-to-byte models," Tech. Rep.

[88] M. Chen, M. D'Arcy, A. Liu, J. Fernandez, and D. Downey, "CODAH: An adversarially-authored question answering dataset for common sense," in *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*. Minneapolis, USA: Association for Computational Linguistics, Jun. 2019, pp. 63–69.

[89] B. Y. Lin, S. Lee, X. Qiao, and X. Ren, "Common Sense Beyond English: Evaluating and Improving Multilingual Language Models for Commonsense Reasoning," Tech. Rep.

[90] J. Yan, M. Raman, A. Chan, T. Zhang, R. Rossi, H. Zhao, S. Kim, N. Lipka, and X. Ren, "Learning Contextualized Knowledge Structures for Commonsense Reasoning," Tech. Rep.

[91] M. Bartolo, A. Roberts, J. Welbl, S. Riedel, and P. Stenetorp, "Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension," Tech. Rep.

[92] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2369–2380.

[93] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," Tech. Rep.

[94] J. Ainslie, S. Ontã nón, C. Alberti, V. Cvicek, Z. Fisher, P. Pham, A. Ravula, S. Sanghai, Q. Wang, and L. Yang Google Research, "ETC: Encoding Long and Structured Inputs in Transformers," Tech. Rep.

[95] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy, and Allen, "SpanBERT: Improving Pre-training by Representing and Predicting Spans," Tech. Rep.

[96] A. Trischler, T. Wang, X. E. Yuan, J. D. Harris, A. Sordoni, P. Bachman, and K. Suleman, "Newsqa: A machine comprehension dataset," November 2016. [Online]. Available: https://www.microsoft.com/en-us/research/publication/newsqa-machine-comprehension-dataset/

[97] W. He, K. Liu, J. Liu, Y. Lyu, S. Zhao, X. Xiao, Y. Liu, Y. Wang, H. Wu, Q. She, X. Liu, T. Wu, and H. Wang, "DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications," Tech. Rep.

[98] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547*, 2020.

[99] A. R. Jafari, G. Li, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "Fine-grained emotions influence on implicit hate speech detection," *IEEE Access*, vol. 11, pp. 105 330–105 343, 2023.

[100] T. Sosea and C. Caragea, "emlm: a new pre-training objective for emotion related tasks," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, pp. 286–293.

[101] M. Hosseini and C. Caragea, "Distilling knowledge for empathy detection," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 3713–3724.

[102] A. R. Jafari, P. Rajapaksha, R. Farahbakhsh, G. Li, and N. Crespi, "Fine-grained emotion knowledge extraction in human values: An interdisciplinary analysis," in *The 12th International Conference on Complex Networks and their Applications*, 2023.

[103] automateddreams, "Natural Language Processing and Why It's So Important — Automated Dreams," 2021, Accessed: 2021-08-03. [Online]. Available: https://www.automateddreams.com/dream-journal/natural-language-processing-and-why-its-so-important

[104] Y. Kuratov and M. Arkhipov, "Adaptation of deep bidirectional multilingual transformers for Russian language," *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, vol. 2019-May, no. 18, pp. 333–339, 2019.

[105] D. Kang and E. Hovy, "xSLUE: A Benchmark and Analysis Platform for Cross-Style Language Understanding and Evaluation," 2019.

[106] L. Stappen, F. Brunn, and B. Schuller, "Cross-lingual Zero- and Few-shot Hate Speech Detection Utilising Frozen Transformer Language Models and AXEL," 2020.

[107] S. K. Tawalbeh, M. Hammad, and M. AL-Smadi, "KEIS@JUST at SemEval-2020 Task 12: Identifying Multilingual Offensive Tweets Using Weighted Ensemble and Fine-Tuned BERT," 2020.

[108] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond English-Centric Multilingual Machine Translation," pp. 1–38.

[109] S. M. Lakew, M. Cettolo, and M. Federico, "A comparison of transformer and recurrent neural networks on multilingual neural machine translation," jun 2018.

[110] J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," pp. 7304–7308, oct 2013.

[111] A. Mohan and R. Rose, "Multi-lingual speech recognition with low-rank multi-task deep neural networks," vol. 2015-August, pp. 4994–4998, aug 2015.

[112] S. Zhou, Y. Zhao, S. Xu, and B. Xu, "Multilingual recurrent neural networks with residual learning for low-resource speech recognition," vol. 2017-August, pp. 704–708, 2017.

[113] S. Zhou, S. Xu, and B. Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," jun 2018.

[114] E. F. Can, A. Ezen-Can, and F. Can, "Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data," 2018.

[115] A. R. Jafari, R. Farahbakhsh, M. Salehi, and N. Crespi, "Language-agnostic method for sentiment analysis of twitter," in *International Conference on Data Analytics & Management*. Springer, 2023, pp. 597–606.

[116] H. Nankani, H. Dutta, H. Shrivastava, P. V. N. S. Rama Krishna, D. Mahata, and R. R. Shah, "Multilingual Sentiment Analysis," 2020, pp. 193–236.

[117] K. Kanclerz, P. Milkowski, and J. Kocon, "Cross-lingual deep neural transfer learning in sentiment analysis," *Procedia Computer Science*, vol. 176, pp. 128–137, 2020.

[118] L. Huang, A. Sil, H. Ji, and R. Florian, "Improving slot filling performance with attentive neural networks on dependency structures," pp. 2588–2597, 2017.

[119] H. Weld, X. Huang, S. Long, J. Poon, and S. C. Han, "A survey of joint intent detection and slot-filling models in natural language understanding," jan 2021.

[120] G. Castellucci, V. Bellomaria, A. Favalli, and R. Romagnoli, "Multilingual Intent Detection and Slot Filling in a Joint BERT-based Model," no. Id, 2019.

[121] J. Luoma and S. Pyysalo, "Exploring Cross-sentence Contexts for Named Entity Recognition with BERT," *arXiv*, 2020.

[122] Q. Wu, Z. Lin, B. Karlsson, J.-G. LOU, and B. Huang, "Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language," no. 2017, pp. 6505–6514, 2020.

[123] M. Moran and C. Lignos, "Effective Architectures for Low Resource Multilingual Named Entity Transliteration," *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pp. 79–86, 2020.

[124] E. Loginova, S. Varanasi, and G. Neumann, "Towards end-to-end multilingual question answering," *Information Systems Frontiers*, vol. 23, no. 1, pp. 227–241, 2021.

[125] J. Clark, "Google AI Blog: TyDi QA: A Multilingual Question Answering Benchmark," 2020. [Online]. Available: https://ai.googleblog.com/2020/02/tydi-qa-multilingual-question-answering.html

[126] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, nov 2017.