A Survey of Techniques for Low-Resource Natural Language Processing

www.surveyx.cn

Abstract

Low-resource natural language processing (NLP) is pivotal in addressing the challenges of languages with limited annotated data. This survey explores the significance, challenges, and innovative methodologies in low-resource NLP, focusing on techniques like few-shot learning, data augmentation, and cross-lingual transfer. Few-shot and zero-shot learning paradigms are essential in environments with scarce data, enabling models to generalize from minimal examples. Data augmentation, through synthetic data generation and self-training, enhances model robustness by increasing training data variability. Cross-lingual transfer techniques leverage high-resource languages to improve low-resource language processing, supported by multilingual pretrained models and initiatives like No Language Left Behind (NLLB). Language model adaptation remains challenging due to data scarcity, requiring innovative strategies for rapid adaptation and knowledge distillation. The survey highlights the importance of efficient data utilization, methodological advancements, and the role of multilingual embeddings in enhancing NLP capabilities across diverse linguistic landscapes. Future research should focus on refining these methodologies, exploring additional NLP tasks, and addressing methodological biases to expand the applicability and effectiveness of NLP technologies in underserved languages.

1 Introduction

1.1 Significance of Low-Resource NLP

Low-resource natural language processing (NLP) is essential for addressing the challenges posed by languages with limited annotated data, exemplified by Ge'ez, where resource scarcity significantly hinders machine translation efforts [1]. The difficulty in extracting structured information from unstructured text in these languages further emphasizes the importance of low-resource NLP [2].

Multilingual Language Models (MLLMs) play a vital role by facilitating the simultaneous processing of multiple languages, thereby improving the understanding of low-resource languages [3]. However, the impracticality of creating comprehensive development sets for truly low-resource languages necessitates the efficient utilization of all available data for training [4].

The limitations of current machine translation systems, which often rely on large volumes of parallel data, highlight the need for few-shot translation approaches for low-resource languages [5]. This need is echoed in conversational benchmarks that reflect real-world scenarios where annotated data is scarce [6].

The significance of low-resource NLP lies in its capacity to enhance communication and comprehension across diverse languages, fostering inclusive technological advancements. Many low-resource languages, such as Sumerian cuneiform, struggle with the absence of pretrained models and development sets. Recent innovations, including cross-lingual information extraction pipelines and text augmentation techniques, have shown promise in improving performance in low-resource contexts.

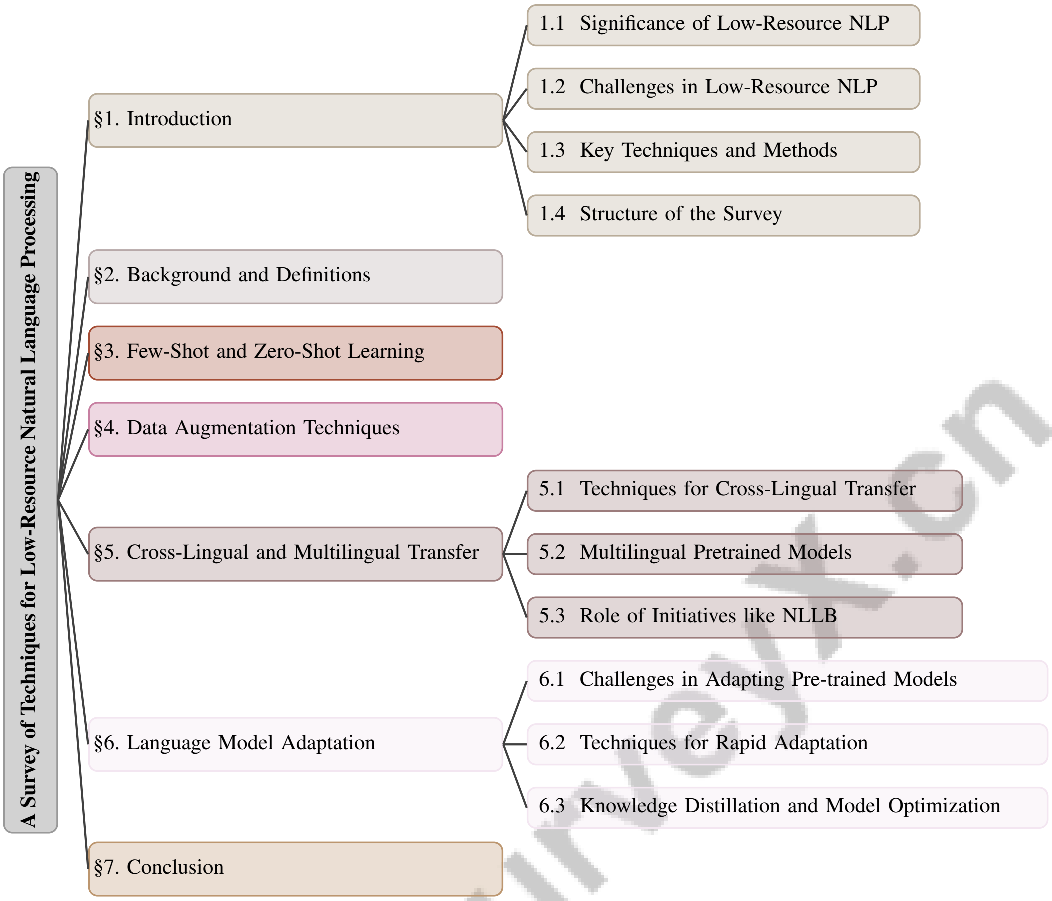

A Survey of Techniques for Low-Resource Natural Language Processing
§1. Introduction
1.1 Significance of Low-Resource NLP
1.2 Challenges in Low-Resource NLP
1.3 Key Techniques and Methods
1.4 Structure of the Survey
§2. Background and Definitions
§3. Few-Shot and Zero-Shot Learning
§4. Data Augmentation Techniques
§5. Cross-Lingual and Multilingual Transfer
5.1 Techniques for Cross-Lingual Transfer
5.2 Multilingual Pretrained Models
5.3 Role of Initiatives like NLLB
§6. Language Model Adaptation
6.1 Challenges in Adapting Pre-trained Models
6.2 Techniques for Rapid Adaptation
6.3 Knowledge Distillation and Model Optimization
§7. Conclusion

Figure 1: chapter structure

By addressing these challenges, low-resource NLP bridges gaps in multilingual natural language understanding and promotes a more equitable digital landscape, enabling diverse linguistic communities to leverage advanced technological tools [7, 8, 4].

1.2 Challenges in Low-Resource NLP

Low-resource NLP faces numerous challenges, primarily due to the scarcity of annotated data and limited resources. The insufficient labeled data presents a critical barrier for training robust models, compounded by the reliance on task-specific models that require significant labeled data often unavailable [2]. Additionally, handling out-of-vocabulary words and domain mismatches complicates matters due to the limited availability of labeled training data [1].

Benchmarks in low-resource NLP frequently assume the existence of development sets, leading to unrealistic performance expectations in real-world scenarios where such resources are scarce [4]. Existing benchmarks often demand extensive training data and fine-tuning, which are costly and time-consuming [6]. These benchmarks primarily focus on task performance without addressing practical aspects like data efficiency, computational costs, and inference latency, which are critical for real-world applications [9].

The performance of large language models (LLMs) in low-data regimes, where only limited task-specific data is available for fine-tuning, poses another significant challenge, often resulting in suboptimal performance [10]. Furthermore, BERT-style pretrained language models (PLMs) show sensitivity to various prompt template designs and discrepancies in word distribution between prompt-style texts and pre-training corpora, creating additional challenges [11].

Transferring knowledge from high-resource to low-resource languages is complicated by methodological biases and potential quality degradation when high-resource datasets are down-sampled [12]. In

AI-generated, for reference only.

multilingual contexts, capacity dilution in MLLMs and the challenge of achieving high performance across diverse languages highlight existing benchmarks' limitations in effectively evaluating MLLM capabilities [3].

Moreover, the reliance on attribute-annotated data for training, which is often unavailable for many lower-resource languages, poses significant challenges for applying existing methods effectively [13]. In specialized domains, such as clinical environments, the assumption of access to large labeled datasets for adaptation is impractical, particularly for rare diseases [14]. These challenges underscore the necessity for innovative approaches to overcome the inherent limitations in low-resource NLP, especially in specialized domains constrained by the scarcity of annotated datasets.

1.3 Key Techniques and Methods

A diverse array of strategies has been developed in low-resource NLP to mitigate the limitations of scarce annotated data. Data augmentation is a cornerstone technique, particularly back-translation, which effectively enhances machine translation systems [15]. The integration of frameworks like AUG-FedPrompt illustrates the application of prompt learning to address few-shot federated NLP tasks [16].

Few-shot and zero-shot learning paradigms are indispensable in low-resource environments. The benchmark introduced by [17] evaluates zero-shot transfer learning alongside unsupervised machine translation, providing a comprehensive assessment of these methodologies. Input reformulation techniques, such as POSE, ParSE, and MiPS, are designed to enhance model conditioning and performance in low-resource translation tasks [18].

Significant advancements in cross-lingual transfer techniques have been made through initiatives like No Language Left Behind (NLLB), which introduced the Flores-200 dataset, a many-to-many multilingual dataset that broadens language coverage and supports cross-lingual NLP tasks [19]. Domain-specific cross-lingual embeddings, as explored by [20], further augment processing capabilities for languages with limited resources by constructing effective seed dictionaries.

The LYRA methodology exemplifies integrated approaches in low-resource settings, combining open LLM fine-tuning, retrieval-augmented generation, and transfer learning to enhance translation quality for rare languages [21]. Additionally, the strategic selection of optimal demonstrations from annotated data, as proposed by [22], highlights the use of smaller models to improve larger LLMs' performance.

Teacher models, such as TeacherLM-7.1B, provide valuable insights by annotating data with fundamentals, chains of thought, and common mistakes, enriching the training process for NLP tasks [23]. Integrating end-task objectives into the training process, as proposed by [24], allows for the simultaneous optimization of both auxiliary and end-tasks, enhancing the efficacy of low-resource NLP models.

Furthermore, the Chain-of-Dictionary Prompting (COD) framework utilizes chains of multilingual dictionaries to improve large language models' translation capabilities [25]. The LLM2LLM framework proposes a targeted, iterative data augmentation approach that employs a teacher LLM to generate synthetic data based on the errors made by a student LLM during fine-tuning [10].

Collectively, these methodologies enhance low-resource NLP by addressing the dual challenges of data scarcity and quality, providing more effective solutions for developing robust NLP models. By critically examining the limitations of down-sampling high-resource language data for low-resource tasks, such as part-of-speech tagging and machine translation, these approaches reveal how data characteristics can significantly influence model performance. The emphasis on data augmentation techniques highlights their potential to enrich training datasets, facilitating the application of NLP models across diverse linguistic landscapes and inspiring further research to optimize these strategies [12, 26].

1.4 Structure of the Survey

This survey is meticulously structured to provide a comprehensive examination of techniques and methodologies pertinent to low-resource natural language processing (NLP). It begins with an introduction that establishes the significance and challenges of low-resource NLP, setting the stage for an in-depth exploration of various strategies employed to address these challenges. Following the introduction is a detailed background section defining core concepts such as few-shot learning, data augmentation, unsupervised alignment, and language model adaptation.

Subsequent sections delve into specific methodologies, with Section 3 focusing on few-shot and zero-shot learning approaches, highlighting their applications and limitations. The survey progresses to a discussion on data augmentation techniques in Section 4, emphasizing the creation of synthetic corpora and self-training methods. Section 5 examines cross-lingual and multilingual transfer strategies, including the role of multilingual embeddings and initiatives like No Language Left Behind (NLLB), while considering frameworks like TransLLM's efficacy in maintaining original knowledge during translation processes [27].

The exploration continues with Section 6, which addresses the adaptation of pre-trained language models to low-resource languages, identifying challenges and discussing innovative adaptation techniques. The survey concludes with a synthesis of key findings and insights, offering perspectives on future research directions and potential advancements in low-resource NLP. This structured approach provides a comprehensive overview of current challenges and advancements in low-resource NLP, highlighting the limitations of existing models, the significance of realistic experimental setups, and innovative techniques such as data augmentation and interpretability tools, while outlining future research directions and potential applications across various low-resource languages [26, 7, 8, 4, 28].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Background and Definitions

Low-resource natural language processing (NLP) contends with substantial obstacles due to the scarcity of annotated data required for training effective machine learning models [29]. This challenge is particularly acute in languages such as Tigrinya, where the lack of parallel data complicates neural machine translation [15]. Addressing these challenges is central to low-resource NLP, which focuses on translating and processing under-resourced languages [25].

Few-shot learning emerges as a pivotal approach, enabling models to generalize from limited labeled examples, a necessity in tasks like named entity recognition (NER) where data is scarce [30]. The constraints of deep learning methods in generalizing from minimal data underscore the importance of few-shot learning [30].

Data augmentation is a critical strategy to mitigate data scarcity by diversifying training examples without new data collection. Techniques such as back-translation and noise injection produce synthetic corpora, enhancing model training in low-resource settings [15]. These methods are crucial for enriching datasets available for low-resource languages [29].

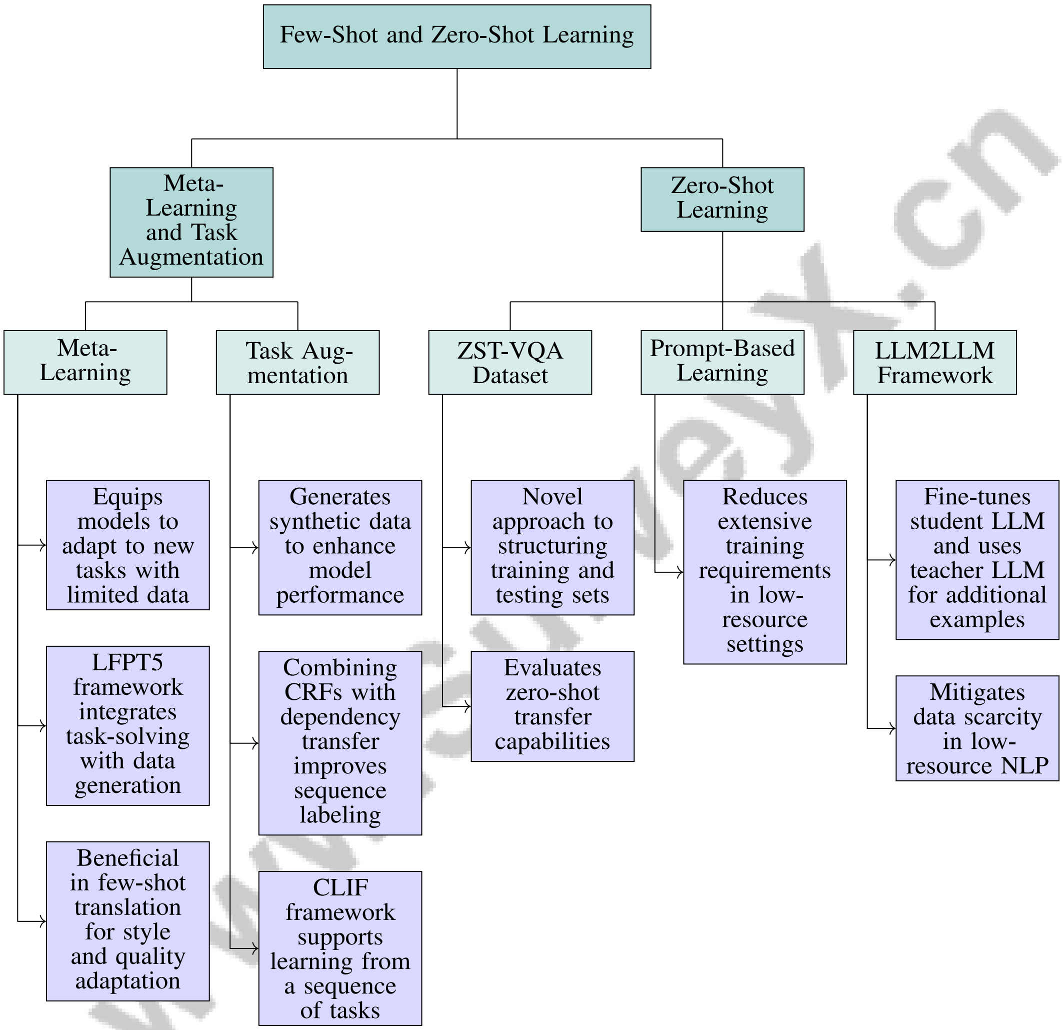
Unsupervised alignment, including sequence segmentation, is vital for languages lacking whitespace-delimited orthography and gold-standard data. This process aligns data like sentences without relying on labeled examples, aiding in the processing of resource-limited languages [17].

Language model adaptation focuses on fine-tuning pre-trained models to specific tasks or domains, yet often struggles due to the need for large datasets, typically scarce in low-resource languages [13]. Effective adaptation enhances model performance across diverse linguistic environments, especially in low-resource contexts [13].

In multilingual scenarios, developing systems that perform NLP tasks across languages while managing limited annotation budgets is a significant challenge. Benchmarks for multilingual instruction-following, particularly those involving translation across languages, require semantic alignment and effective in-context learning by large language models (LLMs) [29]. Multitask learning, which models multiple language generation tasks simultaneously, is employed to boost efficiency and performance in low-resource NLP [24].

These methodologies highlight the complexities and innovative strategies in low-resource NLP, aiming to address challenges posed by data scarcity and linguistic diversity. The field evolves with advanced techniques to overcome low-resource settings' inherent limitations [25].

In recent years, the exploration of learning paradigms has gained significant traction, particularly in the realms of few-shot and zero-shot learning. These strategies not only enhance model performance but also address the challenges associated with limited labeled data. Figure 2 illustrates the hierarchical structure of few-shot and zero-shot learning strategies, highlighting the roles of meta-learning and task augmentation in few-shot learning, along with the contributions of the ZST-VQA dataset, prompt-based learning, and the LLM2LLM framework in zero-shot learning. This visual representation serves to clarify the interconnections between these methodologies, thereby enriching our understanding of their respective contributions to the broader field of machine learning.



Few-Shot and Zero-Shot Learning
Meta-Learning and Task Augmentation
Zero-Shot Learning
Meta-Learning
Task Augmentation
ZST-VQA Dataset
Prompt-Based Learning
LLM2LLM Framework
Equips models to adapt to new tasks with limited data
LFPT5 framework integrates task-solving with data generation
Beneficial in few-shot translation for style and quality adaptation
Generates synthetic data to enhance model performance
Combining CRFs with dependency transfer improves sequence labeling
CLIF framework supports learning from a sequence of tasks
Novel approach to structuring training and testing sets
Evaluates zero-shot transfer capabilities
Reduces extensive training requirements in low-resource settings
Fine-tunes student LLM and uses teacher LLM for additional examples
Mitigates data scarcity in low-resource NLP

Figure 2: This figure illustrates the hierarchical structure of few-shot and zero-shot learning strategies, highlighting the roles of meta-learning and task augmentation in few-shot learning, and the contributions of the ZST-VQA dataset, prompt-based learning, and the LLM2LLM framework in zero-shot learning.

3 Few-Shot and Zero-Shot Learning

3.1 Meta-Learning and Task Augmentation

Meta-learning and task augmentation are pivotal in advancing few-shot and zero-shot learning within low-resource natural language processing (NLP). Meta-learning, often described as "learning to learn," equips models to swiftly adapt to new tasks with limited data, effectively addressing the scarcity of annotated resources. The LFPT5 framework exemplifies this by integrating task-solving

AI-generated, for reference only.

with data generation, emulating the human ability to learn from minimal examples [31]. This method is particularly beneficial in few-shot translation, enabling models to adjust to the style and quality of demonstrations, thereby producing high-quality outputs despite limited training data [5].

Task augmentation complements meta-learning by generating synthetic data to enhance model performance across various NLP tasks. For example, combining conditional random fields (CRFs) with dependency transfer mechanisms significantly improves sequence labeling in few-shot contexts [32]. The Continual Learning of Few-Shot Learners (CLIF) framework further supports models in learning from a sequence of tasks while maintaining the ability to generalize to new ones [33].

In zero-shot learning, the ZST-VQA dataset introduces a novel approach to structuring training and testing sets, effectively evaluating zero-shot transfer capabilities and addressing deficiencies in existing benchmarks [34]. Prompt-based learning methods enhance this by reducing extensive training requirements, which is particularly advantageous in low-resource settings [6].

The LLM2LLM framework illustrates the potential of task augmentation by fine-tuning a student large language model (LLM), evaluating its performance on seed data, and using a teacher LLM to generate additional examples based on identified errors [10].

As illustrated in Figure 3, the hierarchical structure of Meta-Learning and Task Augmentation in low-resource NLP highlights key frameworks and datasets. This figure categorizes the main approaches into Meta-Learning, Task Augmentation, and Zero-Shot Learning, each associated with specific methods and datasets that enhance model performance in few-shot and zero-shot scenarios. Collectively, these strategies mitigate data scarcity in low-resource NLP, improving the applicability and performance of models across diverse linguistic environments.

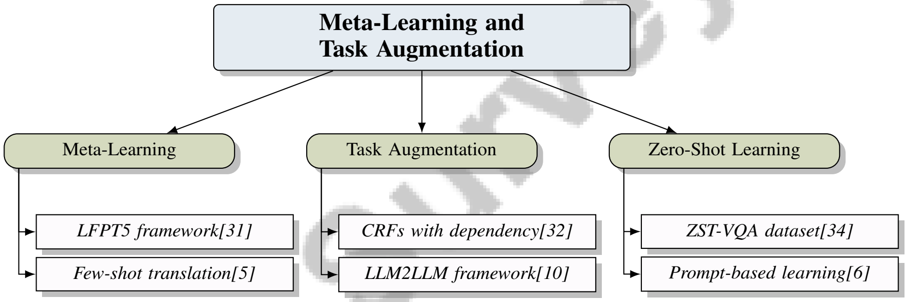


Figure 3: This figure illustrates the hierarchical structure of Meta-Learning and Task Augmentation in low-resource NLP, highlighting key frameworks and datasets. It categorizes the main approaches into Meta-Learning, Task Augmentation, and Zero-Shot Learning, each associated with specific methods and datasets that enhance model performance in few-shot and zero-shot scenarios.

4 Data Augmentation Techniques

Category	Feature	Method
Synthetic Data Generation and Self-Training	Synthetic Data Techniques Low-Resource Strategies	BiHNet+Reg[33] BT[15], COD[25], FST[5], FS-LDTP[32]
Continuous Semantic and Text-Based Augmentation	Task-Oriented Training Synthetic Data Generation	KnowDA[35] LLM2LLM[10], LFPT5[31]

Table 1: This table presents a comprehensive overview of various data augmentation techniques categorized under Synthetic Data Generation and Self-Training, and Continuous Semantic and Text-Based Augmentation. It highlights key features and methods employed in these categories, referencing prominent frameworks and their contributions to enhancing model performance in low-resource natural language processing (NLP) environments.

Data augmentation techniques play a crucial role in overcoming challenges associated with low-resource NLP environments by amplifying the quantity and enhancing the diversity of training data. Strategies such as context manipulation and text simplification introduce varied writing styles, thereby creating enriched training examples that significantly improve neural language model performance in these settings. Additionally, ensemble learning, which leverages predictions from multiple pretrained models, aids in effective model training, particularly in specialized fields with limited annotated data [36, 37]. Synthetic data generation emerges as a prominent strategy, producing additional training instances that bolster model performance. Table 1 provides a detailed categorization of data augmentation techniques, focusing on synthetic data generation, self-training, and continuous semantic and text-based augmentation methods, crucial for improving model performance in low-resource NLP contexts. This section delves into techniques like Synthetic Data Generation and Self-Training, vital for augmenting training datasets in low-resource contexts.

4.1 Synthetic Data Generation and Self-Training

Synthetic data generation and self-training are pivotal methodologies for enhancing model performance in low-resource NLP environments. By increasing the variability of training instances, synthetic data generation improves model robustness. Techniques like back-translation generate synthetic parallel data, enhancing translations for low-resource languages [15]. The Chain-of-Dictionary Prompting (COD) framework exemplifies this by augmenting translation prompts with multilingual dictionary chains, offering contextual translations for specific words [25].

The LFPT5 framework demonstrates the use of pseudo samples from prior tasks to aid knowledge retention and minimize forgetting, essential for continuous learning in resource-constrained settings [31]. Few-shot translation methods utilizing a transformer decoder-only model trained with self-supervised learning show effectiveness in translating based on limited examples, addressing data scarcity challenges [5].

Self-training, which employs a model's predictions as pseudo-labels for unlabeled data, is another critical technique. The LLM2LLM framework iteratively enhances a small seed dataset with synthetic examples derived from a student large language model's incorrect predictions [10]. This iterative process enriches the dataset and improves model performance in low-resource scenarios.

Furthermore, integrating token similarities and label dependencies in few-shot sequence labeling tasks showcases the potential of synthetic data generation to augment training datasets [32]. The continual learning framework by [33] facilitates few-shot adaptation by generating adapter weights for a frozen BART model, supporting continuous learning and mitigating forgetting.

Collectively, these methodologies enhance low-resource NLP by employing innovative strategies such as data augmentation and neural ensemble learning to effectively tackle data scarcity, particularly in specialized domains and underrepresented languages. By generating diverse training data and optimizing model predictions, these approaches significantly improve neural language models' performance across various NLP tasks, addressing both the quantity and quality of data needed for effective model training [12, 37, 26]. Through synthetic data generation and self-training, these strategies offer robust solutions to challenges encountered in low-resource environments.

4.2 Continuous Semantic and Text-Based Augmentation

Method Name	Data Augmentation	Learning Strategies	Resource Efficiency
KnowDA[35]	Diverse Synthetic Data	Multi-task Training	Minimal Supervision
LLM2LLM[10]	Targeted Augmentation	Iterative Data Augmentation	Small Seed Dataset
LFPT5[31]	Pseudo Samples	Prompt Tuning	Few-shot Tasks

Table 2: Comparison of continuous semantic and text-based augmentation methods in low-resource NLP scenarios, focusing on data augmentation, learning strategies, and resource efficiency. The table includes the KnowDA, LLM2LLM, and LFPT5 frameworks, highlighting their unique approaches to synthetic data generation, training methodologies, and minimal supervision requirements.

Continuous semantic and text-based augmentation techniques are essential for enriching data variability in low-resource NLP scenarios. The KnowDA framework generates diverse synthetic data through a multi-task training approach and a novel auto-regressive generation framework, significantly enhancing the contextual relevance of training datasets [35]. This approach is particularly beneficial in scenarios with limited annotated data, enabling models to learn from a richer set of examples.

The LLM2LLM framework further demonstrates the effectiveness of targeted augmentation, focusing on challenging examples to foster efficient learning and improve performance in low-data scenarios

AI-generated, for reference only.

[10]. This targeted strategy ensures that augmented data is diverse and strategically enhances the model's capability to handle difficult cases, thus improving overall robustness.

Incorporating pseudo samples into training, as shown in the LFPT5 framework, involves generating synthetic data from previously learned tasks and blending it with new task data. This method facilitates continuous learning and adaptation by tuning prompt embeddings for new task types, enhancing the model's ability to generalize across various tasks and domains [31].

Benchmark protocols introduced by [38] emphasize using few labeled examples alongside unsupervised criteria for model selection, highlighting the potential of minimal supervision to achieve effective model performance, especially in low-resource settings where labeled data is scarce.

These continuous semantic and text-based augmentation techniques collectively advance low-resource NLP by enhancing data variability and improving model robustness. Methodologies that generate diverse and contextually relevant synthetic data, such as the "generate, annotate, and learn" (GAL) framework, effectively address data scarcity challenges in NLP. Strategies like knowledge distillation, self-training, and few-shot learning enhance NLP model training. While synthetic datasets can reliably benchmark simpler tasks like intent classification, their effectiveness diminishes for complex tasks such as named entity recognition. To mitigate biases from using the same large language models for both data generation and task performance, it is advisable to utilize data generated from multiple larger models. This comprehensive approach facilitates effective learning and broadens the applicability of NLP models across various linguistic contexts [39, 40]. Table 2 presents a comparative analysis of various frameworks for continuous semantic and text-based augmentation, emphasizing their contributions to addressing data scarcity in low-resource natural language processing (NLP) environments.

Feature	**Synthetic Data Generation and Self-Training**	**Continuous Semantic and Text-Based Augmentation**
Core Strategy	Data Synthesis	Semantic Augmentation
Primary Benefit	Model Robustness	Data Variability
Application Context	Low-resource Nlp	Low-resource Scenarios

Table 3: Comparison of data augmentation techniques focusing on synthetic data generation and self-training versus continuous semantic and text-based augmentation. The table highlights core strategies, primary benefits, and application contexts, illustrating their roles in enhancing model performance in low-resource NLP environments.

5 Cross-Lingual and Multilingual Transfer

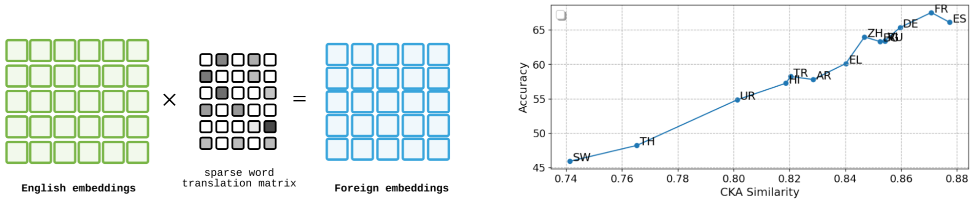
5.1 Techniques for Cross-Lingual Transfer

Cross-lingual transfer techniques extend NLP capabilities to low-resource languages by leveraging data from high-resource languages. The Multilingual Neural Machine Translation (MNMT) model exemplifies this by using linguistic similarities among related languages, such as Ge'ez, to improve translation quality in resource-constrained settings [1]. Multilingual pre-trained language models (mPLMs), including mBART and NLLB-200, demonstrate effectiveness in cross-lingual transfer through benchmark analyses, enhancing applicability across diverse languages [41]. Translation pair prediction (TPP) further bolsters mBERT's zero-shot multilingual transfer capabilities [42].

Challenges in achieving language-neutral representations persist. mBERT's limitations in semantic transfer highlight the need for improved language-neutral representations [43]. Few-shot learning enhances translation quality with minimal resources, allowing models to adapt to new languages with limited examples [5]. These techniques address significant challenges in cross-lingual NLP, such as leveraging pre-trained models for effective zero-shot transfer and employing data augmentation to improve model performance.

Approaches using unlabelled text for language detection and innovative data augmentation significantly extend language model capabilities, fostering equitable growth across diverse linguistic landscapes [44, 3, 26, 45]. By strategically employing bilingual data and shared linguistic features, these methodologies enhance cross-lingual transfer effectiveness.

As shown in Figure 4, various techniques facilitate linguistic feature transfer across languages. The "Sparse Word Translation Matrix" illustrates a method for translating sparse words using a translation



(a) Sparse Word Translation Matrix[45]

(b) The graph shows the relationship between the accuracy of a model and the similarity of its predictions to the ground truth labels for different languages.[46]

Figure 4: Examples of Techniques for Cross-Lingual Transfer

matrix that connects English embeddings with foreign language embeddings. The "Accuracy vs. CKA Similarity" graph highlights the correlation between model accuracy and prediction similarity across languages, emphasizing prediction similarity's significance for achieving high accuracy in multilingual contexts [45, 46].

5.2 Multilingual Pretrained Models

Multilingual pretrained models (MMPLMs) advance cross-lingual transfer by utilizing linguistic knowledge from diverse languages, thereby enhancing NLP capabilities. Models such as mT5, mBART, and NLLB-200 effectively capture syntactic and semantic nuances, enabling zero-shot transfer tasks [47]. Frameworks like the one proposed by [48] use transferable phoneme embeddings to create a unified latent space for phonemes from various languages, facilitating better adaptation and generalization.

The Subword Mapping and Anchoring across Languages (SMALA) method improves bilingual subword vocabularies, enhancing subword alignment crucial for multilingual model performance [49]. A centering procedure introduced by [43] enhances language neutrality within mBERT representations, essential for effective cross-lingual transfer. Supervised approaches significantly outperform in-context learning in multilingual NLP tasks, highlighting the need for refined multilingual pretraining strategies [9].

MMPLMs, such as mBERT and XLM-R, are foundational in cross-lingual NLP, enabling zero-shot transfer learning across numerous languages. Research efforts focus on building larger multilingual models, creating diverse evaluation benchmarks, and enhancing performance on both seen and unseen languages. Innovative model selection techniques demonstrate that superior models can be identified even with limited annotated data in auxiliary languages, thus improving cross-lingual capabilities across various tasks [50, 3, 45]. These models broaden the applicability of NLP systems by leveraging shared linguistic features and enhancing language neutrality.

5.3 Role of Initiatives like NLLB

Initiatives such as No Language Left Behind (NLLB) significantly advance cross-lingual transfer by addressing low-resource language challenges and promoting digital inclusivity. NLLB has developed comprehensive multilingual datasets, like MultiEURLEX, facilitating cross-lingual classification research [51]. This initiative underscores potential improvements in translating low-resource languages [19].

Advanced multilingual models, including NLLB-200, highlight NLLB initiatives' contributions to enhancing cross-lingual transfer and language technology accessibility [13]. Despite challenges, such as GPT-4's superior performance in certain translation tasks, NLLB remains a critical benchmark for evaluating multilingual machine translation capabilities [29]. The Chain-of-Dictionary Prompting (COD) framework surpasses the state-of-the-art NLLB 3.3B translator in many instances, illustrating NLLB's role in promoting effective cross-lingual transfer [25].

Future research areas include integrating additional linguistic features and refining difficulty measurement processes to enhance cross-lingual transfer and code-switching data generation [47]. These initiatives collectively advance cross-lingual NLP by fostering inclusive and scalable language technologies, thereby improving the applicability of NLP models across diverse linguistic landscapes.

6 Language Model Adaptation

6.1 Challenges in Adapting Pre-trained Models

Adapting pre-trained models to low-resource languages is fraught with challenges, primarily due to the lack of annotated data and the linguistic diversity of these languages. A major issue is the overfitting of models to the source language during fine-tuning, which diminishes their generalizability in low-resource contexts [41]. This problem is compounded by the inconsistent performance of Multilingual Language Models (MLLMs) across languages with limited resources [3].

The concurrent classification of new and fixed classes remains a challenge for existing few-shot learning methods [52]. The robustness of language-neutral components in models like mBERT is insufficient for tasks such as machine translation quality estimation [43]. Studies indicate that pre-trained models do not always outperform meta-learning models, with dataset diversity significantly affecting performance [53].

Adaptation is further complicated by reliance on large development sets for early stopping, often leading to overestimated model performance in low-resource NLP research [4]. Limited data for new tasks can exacerbate overfitting, resulting in the forgetting of previously learned knowledge [31]. Catastrophic forgetting, where models lose performance on earlier tasks when new ones are introduced, remains unresolved [33].

The risk of data contamination and limited shots for few-shot learning further restrict these methods [6]. These challenges necessitate innovative strategies to enhance pre-trained model adaptability to low-resource languages, focusing on bias reduction, training transparency, and robust data augmentation techniques.

6.2 Techniques for Rapid Adaptation

Rapid adaptation of language models to new tasks and domains is crucial in low-resource NLP environments, where data scarcity is a significant challenge. Innovative techniques focus on parameter optimization and enhanced learning capabilities. The LFPT5 framework exemplifies this by retaining knowledge through pseudo sample generation and minimizing forgetting via prompt embedding tuning, facilitating continuous learning and rapid adaptation [31].

The LLM2LLM framework enhances performance in low-data regimes by iteratively augmenting small seed datasets with synthetic examples from student LLM errors, significantly increasing accuracy across datasets [10]. This method not only enriches datasets but also strengthens model robustness in low-resource scenarios.

Unified Prompt Tuning (UPT) improves pre-trained language models' (PLMs) generalization by learning from diverse, non-target task groups, enhancing performance in few-shot learning scenarios. UPT captures prompting semantics across NLP tasks, allowing models to acquire task-invariant knowledge and adapt effectively to unfamiliar tasks. Incorporating a self-supervised learning task further boosts generalization, yielding superior results in low-resource settings compared to traditional prompt-based fine-tuning methods [54, 55, 11, 2, 33]. Subword mapping techniques like SMALA address cross-lingual adaptation by utilizing subword similarities to create anchors, enhancing multilingual model performance through accurate subword alignment.

Advancements in unsupervised methods that address morphological complexity and leverage weak supervision are crucial for future research, particularly in enhancing rapid adaptation techniques. Meta-learning and unsupervised language models can improve low-resource performance, while methods like Unsupervised Data Augmentation (UDA) show that model prediction consistency can be achieved without complex data augmentation or reliance on unlabeled data. The transfer of unsupervised sequence segmentation techniques to low-resource languages suggests significant improvements through multilingual pretraining, especially in data-limited situations [38, 56, 57, 58, 59]. Exploring alternative early stopping methods could further enhance model adaptation efficiency in low-resource NLP.

AI-generated, for reference only.

These innovative techniques collectively advance low-resource NLP by providing robust solutions for rapid model adaptation. By aligning embeddings, utilizing synthetic data generation methods like the "generate, annotate, and learn" (GAL) framework, and implementing data augmentation techniques, these methodologies significantly improve language model adaptability and performance across various tasks and domains, particularly in low-resource scenarios. This approach facilitates effective knowledge distillation and self-training, addressing data scarcity challenges and leading to substantial performance gains in applications such as intent classification and dependency parsing [39, 37, 40, 7].

6.3 Knowledge Distillation and Model Optimization

Knowledge distillation and model optimization are crucial for enhancing language model adaptability to low-resource NLP tasks. These methods focus on transferring knowledge from larger, complex models to smaller, efficient ones, improving performance while reducing computational demands. The LetzTranslate approach exemplifies this by producing high-performance bilingual machine translation models with lower computational needs, advantageous in low-resource settings [60].

Knowledge distillation involves a teacher model guiding a smaller student model, transferring learned representations and decision boundaries. This technique compresses model knowledge while maintaining performance, particularly beneficial in low-resource environments with constrained computational capabilities. This is vital in NLP, where pretrained neural models often struggle due to data scarcity, especially in specialized domains like medical applications or underrepresented languages. Methods like data augmentation and neural ensemble learning enhance model adaptability, ensuring reliable results despite limited data and computational resources [37, 8].

Model optimization refines internal representations of language models to better suit target languages. The LMS method uses internal representations and language embeddings to rank models based on target language performance, improving cross-lingual transfer capabilities beyond traditional methods reliant on English validation data [50].

Techniques like soft layer selection in meta-learning frameworks reduce the need for extensive meta-parameters, demonstrating efficiency compared to approaches like X-MAML [61]. This contributes to more efficient layer selection, optimizing model architecture for specific tasks and domains in low-resource scenarios.

These strategies collectively advance low-resource NLP by offering robust solutions to model adaptation challenges. By employing knowledge distillation and model optimization techniques, such as the Generate, Annotate, and Learn (GAL) framework and Knowledge Mixture Data Augmentation (KnowDA), researchers enhance language model efficiency and performance. These methodologies improve capabilities in low-resource linguistic environments, where annotated data is scarce, such as in specialized medical domains or underrepresented languages, enabling the synthesis of high-quality task-specific text. This facilitates broader applications of language models across diverse linguistic landscapes, even when resources are limited [37, 35, 62, 40, 4].

7 Conclusion

7.1 Innovations and Future Directions

Advancements in low-resource natural language processing (NLP) are essential for addressing the challenges of data scarcity and linguistic diversity. Future research should focus on refining inference-time control methods in complex attribute scenarios to enhance model adaptability across various linguistic contexts. Developing robust strategies for generating low-resource datasets and broadening the scope of NLP tasks examined are critical for advancing the field and mitigating methodological biases.

The LFPT5 framework offers a promising direction for future exploration by improving the quality of generated pseudo samples and evaluating its adaptability across diverse tasks, thereby strengthening the robustness of continual learning models. Additionally, exploring the reduction of model size while investigating multilingual capabilities within few-shot learning paradigms could allow for the integration of larger datasets without sacrificing the benefits of minimal data requirements.

AI-generated, for reference only.

In sequence labeling, refining dependency transfer mechanisms and examining additional contextual embeddings could enhance model performance in complex scenarios. Improving continual learning algorithms to integrate seamlessly with large-scale pre-trained models and exploring task-agnostic frameworks may lead to more flexible and efficient learning systems.

Future research could also explore additional dialogue tasks, refine prompt engineering, and evaluate benchmarks through human assessments to enhance the applicability of few-shot learning in conversational AI. By adopting these innovative approaches, research can effectively overcome current limitations in low-resource NLP, thereby expanding the reach and efficacy of NLP technologies across underserved languages and domains.

References

[1] Aman Kassahun Wassie. Machine translation for ge’ez language, 2024.

[2] Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. Zero-shot information extraction as a unified text-to-triple translation, 2021.

[3] Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. A primer on pretrained multilingual language models, 2021.

[4] Katharina Kann, Kyunghyun Cho, and Samuel R. Bowman. Towards realistic practices in low-resource natural language processing: The development set, 2019.

[5] Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. The unreasonable effectiveness of few-shot learning for machine translation, 2023.

[6] Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. Few-shot bot: Prompt-based learning for dialogue systems, 2021.

[7] Gözde Gül Şahin. To augment or not to augment? a comparative study on text augmentation techniques for low-resource nlp, 2021.

[8] Rachit Bansal, Himanshu Choudhary, Ravneet Punia, Niko Schenk, Jacob L Dahl, and Émilie Pagé-Perron. How low is too low? a computational perspective on extremely low-resource languages, 2021.

[9] Evgeniia Razumovskaia, Ivan Vulić, and Anna Korhonen. Analyzing and adapting large language models for few-shot multilingual nlu: Are we there yet?, 2024.

[10] Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipalli, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. Llm2llm: Boosting llms with novel iterative data enhancement, 2024.

[11] Jianing Wang, Chengyu Wang, Fuli Luo, Chuanqi Tan, Minghui Qiu, Fei Yang, Qiuhui Shi, Songfang Huang, and Ming Gao. Towards unified prompt tuning for few-shot text classification, 2022.

[12] Maartje ter Hoeve, David Grangier, and Natalie Schluter. High-resource methodological bias in low-resource investigations, 2022.

[13] Danni Liu and Jan Niehues. How transferable are attribute controllers on pretrained multilingual translation models?, 2024.

[14] Fereshteh Shakeri, Yunshi Huang, Julio Silva-Rodríguez, Houda Bahig, An Tang, Jose Dolz, and Ismail Ben Ayed. Few-shot adaptation of medical vision-language models, 2024.

[15] Lidia Kidane, Sachin Kumar, and Yulia Tsvetkov. An exploration of data augmentation techniques for improving english to tigrinya translation, 2021.

[16] Dongqi Cai, Yaozong Wu, Haitao Yuan, Shangguang Wang, Felix Xiaozhu Lin, and Mengwei Xu. Towards practical few-shot federated nlp, 2023.

[17] Aviral Joshi, Chengzhi Huang, and Har Simrat Singh. Zero-shot language transfer vs iterative back translation for unsupervised machine translation, 2021.

[18] Brian Yu, Hansen Lillemark, and Kurt Keutzer. Simple and effective input reformulations for translation. *arXiv preprint arXiv:2311.06696*, 2023.

[19] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.

[20] Lena Shakurova, Beata Nyari, Chao Li, and Mihai Rotaru. Best practices for learning domain-specific cross-lingual embeddings, 2019.

[21] Ibrahim Merad, Amos Wolf, Ziad Mazzawi, and Yannick Léo. Language very rare for all, 2024.

[22] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning, 2024.

[23] Nan He, Hanyu Lai, Chenyang Zhao, Zirui Cheng, Junting Pan, Ruoyu Qin, Ruofan Lu, Rui Lu, Yunchen Zhang, Gangming Zhao, Zhaohui Hou, Zhiyuan Huang, Shaoqing Lu, Ding Liang, and Mingjie Zhan. Teacherlm: Teaching to fish rather than giving the fish, language modeling likewise, 2024.

[24] Lucio M. Dery, Paul Michel, Ameet Talwalkar, and Graham Neubig. Should we be pre-training? an argument for end-task aware training as an alternative, 2022.

[25] Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. Chain-of-dictionary prompting elicits translation in large language models, 2024.

[26] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp, 2021.

[27] Xiang Geng, Ming Zhu, Jiahuan Li, Zhejian Lai, Wei Zou, Shuaijie She, Jiaxin Guo, Xiaofeng Zhao, Yinglu Li, Yuang Li, et al. Why not transform chat large language models to non-english? *arXiv preprint arXiv:2405.13923*, 2024.

[28] Haoyue Shi, Karen Livescu, and Kevin Gimpel. Substructure substitution: Structured data augmentation for nlp, 2021.

[29] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis, 2024.

[30] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning, 2022.

[31] Chengwei Qin and Shafiq Joty. Lfpt5: A unified framework for lifelong few-shot language learning based on prompt tuning of t5, 2022.

[32] Yutai Hou, Zhihan Zhou, Yijia Liu, Ning Wang, Wanxiang Che, Han Liu, and Ting Liu. Few-shot sequence labeling with label dependency transfer and pair-wise embedding, 2019.

[33] Xisen Jin, Bill Yuchen Lin, Mohammad Rostami, and Xiang Ren. Learn continually, generalize rapidly: Lifelong knowledge accumulation for few-shot learning, 2022.

[34] Yuanpeng Li, Yi Yang, Jianyu Wang, and Wei Xu. Zero-shot transfer vqa dataset, 2018.

[35] Yufei Wang, Jiayi Zheng, Can Xu, Xiubo Geng, Tao Shen, Chongyang Tao, and Daxin Jiang. Knowda: All-in-one knowledge mixture model for data augmentation in low-resource nlp, 2023.

[36] Yujin Kim, Jaehoon Oh, Sungnyun Kim, and Se-Young Yun. How to fine-tune models with few samples: Update, data augmentation, and test-time augmentation, 2022.

[37] Hoang Van. Mitigating data scarcity for large language models, 2023.

[38] Haoyue Shi, Karen Livescu, and Kevin Gimpel. On the role of supervision in unsupervised constituency parsing, 2020.

[39] Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. Efficacy of synthetic data as a benchmark, 2024.

[40] Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. Generate, annotate, and learn: Nlp with synthetic text, 2022.

[41] Nadezhda Chirkova, Sheng Liang, and Vassilina Nikoulina. Empirical study of pretrained multilingual language models for zero-shot cross-lingual knowledge transfer in generation, 2024.

[42] Shubhanshu Mishra and Aria Haghighi. Improved multilingual language model pretraining for social media text via translation pair prediction, 2021.

[43] Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. How language-neutral is multilingual bert?, 2019.

[44] Louis Clouâtre, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar. Detecting languages unintelligible to multilingual models through local structure probes, 2022.

[45] Ke Tran. From english to foreign languages: Transferring pre-trained language models, 2020.

[46] Shanu Kumar, Abbaraju Soujanya, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. Ditto: A feature representation imitation approach for improving cross-lingual transfer, 2023.

[47] Ningyu Xu, Tao Gui, Ruotian Ma, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. Cross-linguistic syntactic difference in multilingual bert: How good is it and how does it affect transfer?, 2022.

[48] Wei-Ping Huang, Po-Chun Chen, Sung-Feng Huang, and Hung yi Lee. Few-shot cross-lingual tts using transferable phoneme embedding, 2022.

[49] Giorgos Vernikos and Andrei Popescu-Belis. Subword mapping and anchoring across languages, 2021.

[50] Yang Chen and Alan Ritter. Model selection for cross-lingual transfer, 2021.

[51] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer, 2021.

[52] Yangbin Chen, Tom Ko, Lifeng Shang, Xiao Chen, Xin Jiang, and Qing Li. An investigation of few-shot learning in spoken term classification, 2020.

[53] Brando Miranda, Patrick Yu, Saumya Goyal, Yu-Xiong Wang, and Sanmi Koyejo. Is pre-training truly better than meta-learning?, 2023.

[54] Robert L. Logan IV au2, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. Cutting down on prompts and parameters: Simple few-shot learning with language models, 2021.

[55] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners, 2021.

[56] C. M. Downey, Shannon Drizin, Levon Haroutunian, and Shivin Thukral. Multilingual unsupervised sequence segmentation transfers to extremely low-resource languages, 2022.

[57] David Lowell, Brian E. Howard, Zachary C. Lipton, and Byron C. Wallace. Unsupervised data augmentation with naive augmentation and without unlabeled data, 2020.

[58] Shumin Deng, Ningyu Zhang, Zhanlin Sun, Jiaoyan Chen, and Huajun Chen. When low resource nlp meets unsupervised language model: Meta-pretraining then meta-learning for few-shot text classification, 2019.

[59] Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction, 2018.

[60] Yewei Song, Saad Ezzini, Jacques Klein, Tegawende Bissyande, Clément Lefebvre, and Anne Goujon. Letz translate: Low-resource machine translation for luxembourgish, 2023.

[61] Weijia Xu, Batool Haider, Jason Krone, and Saab Mansour. Soft layer selection with meta-learning for zero-shot cross-lingual transfer, 2021.

[62] Yukun Huang, Yanda Chen, Zhou Yu, and Kathleen McKeown. In-context learning distillation: Transferring few-shot learning ability of pre-trained language models, 2022.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.