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 of languages with limited annotated data. This is particularly evident in languages such as Ge'ez, where resource scarcity significantly hampers machine translation efforts [1]. The difficulty in extracting structured information from unstructured text in these languages highlights the critical need for low-resource NLP solutions [2].

Multilingual Language Models (MLLMs) play a vital role by facilitating the simultaneous processing of multiple languages, which enhances comprehension of low-resource languages [3]. However, the challenge of developing comprehensive development sets for truly low-resource languages necessitates the efficient use of all available data for training [4]. Furthermore, the limitations of current machine translation systems, which often rely on extensive parallel data, emphasize the importance of few-shot translation approaches for low-resource languages [5]. The ability to engage in meaningful conversations with minimal examples, as evaluated in conversational benchmarks, reflects real-world scenarios where annotated data is scarce [6].

The significance of low-resource NLP extends to enhancing communication across diverse languages, fostering inclusive technological innovations. The challenges of applying advanced deep learning techniques to low-resource languages have led to the development of pioneering tools, including cross-lingual information extraction pipelines and interpretability toolkits. These initiatives address the lack of pretrained models and underscore the importance of realistic experimental setups and text

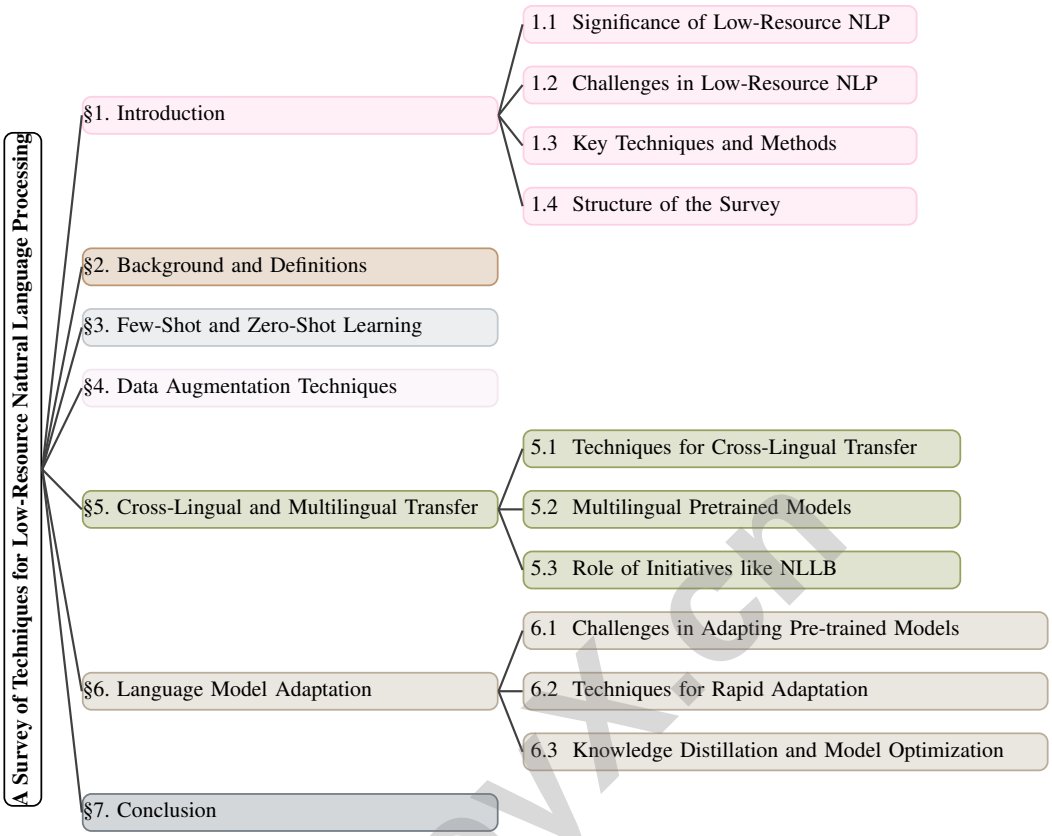


Figure 1: chapter structure

augmentation strategies, which can substantially improve performance in tasks like part-of-speech tagging and dependency parsing. Ultimately, low-resource NLP is crucial for bridging the gap in multilingual natural language understanding, promoting equitable access to technology across various linguistic communities [4, 7, 8].

1.2 Challenges in Low-Resource NLP

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

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

The performance of large language models (LLMs) in low-data regimes, where only minimal task-specific data is available for fine-tuning, presents another challenge, often resulting in suboptimal model performance [10]. Furthermore, BERT-style pretrained language models (PLMs) are sensitive to variations in prompt templates and discrepancies in word distribution between prompt-style texts and pre-training corpora, adding to the complexity [11].

Transferring knowledge from high-resource to low-resource languages is further complicated by methodological biases and potential quality degradation when high-resource datasets are down-sampled [12]. In multilingual contexts, capacity dilution in MLLMs and the challenge of achieving

high performance across diverse languages underscore the limitations of existing benchmarks in evaluating MLLM capabilities effectively [3].

Additionally, 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 settings, the assumption of access to large labeled datasets for adaptation is impractical, particularly for rare diseases [14]. These challenges highlight the need for innovative approaches to overcome inherent limitations in low-resource NLP, especially in specialized domains constrained by the scarcity of annotated datasets.

1.3 Key Techniques and Methods

A variety of strategies have been developed in low-resource NLP to mitigate the limitations imposed by scarce annotated data. Data augmentation is a fundamental technique, with back-translation proving particularly effective in enhancing machine translation systems [15]. The integration of prompt learning through frameworks like AUG-FedPrompt addresses few-shot federated NLP tasks [16].

Few-shot and zero-shot learning paradigms are crucial in low-resource environments. The benchmark established 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, enhance model conditioning and performance in low-resource translation tasks [18].

Cross-lingual transfer techniques have advanced significantly through initiatives like No Language Left Behind (NLLB), which introduced the Flores-200 dataset, a many-to-many multilingual dataset that expands language coverage and supports cross-lingual NLP tasks [19]. Domain-specific cross-lingual embeddings, as explored by [20], further enhance 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 improve translation quality for rare languages [21]. Additionally, the strategic selection of optimal demonstrations from annotated data, as proposed by [22], illustrates the use of smaller models to enhance the performance of larger LLMs.

Teacher models, such as TeacherLM-7.1B, annotate 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 suggested by [24], enables simultaneous optimization of auxiliary and end-tasks, enhancing the efficacy of low-resource NLP models.

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

These methodologies collectively advance low-resource NLP by providing effective solutions to data scarcity challenges, improving data quality through advanced augmentation techniques, and addressing the limitations of down-sampling high-resource language data. This not only enhances the performance of NLP models in low-resource scenarios, particularly in tasks like part-of-speech tagging and machine translation, but also broadens the applicability of these models across diverse linguistic contexts, ultimately paving the way for more equitable advancements in NLP [12, 26].

1.4 Structure of the Survey

This survey is structured to provide a thorough examination of techniques and methodologies relevant to low-resource 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. Section 4 discusses data augmentation techniques, emphasizing synthetic corpus creation 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 also considering the efficacy of frameworks like TransLLM in preserving original knowledge during translation processes [27].

The exploration continues with Section 6, addressing the adaptation of pre-trained language models to low-resource languages, identifying challenges and discussing innovative adaptation techniques. The survey concludes by synthesizing 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 future opportunities in low-resource NLP, particularly emphasizing the limitations of existing deep learning models due to the scarcity of pretrained resources and the necessity for realistic experimental practices, such as the use of development sets, to ensure accurate performance assessments across various languages and tasks [4, 8].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) addresses the challenges of processing languages with insufficient annotated data, crucial for training machine learning models effectively [28]. This challenge is particularly evident in languages like Tigrinya, where the lack of parallel data significantly hinders neural machine translation [15]. The field is dedicated to the translation and processing of such languages [25].

Few-shot learning is integral to low-resource NLP, enabling models to generalize from limited labeled examples. It is especially beneficial for tasks such as named entity recognition (NER), where entity classification with minimal labeled data is challenging [29]. These challenges highlight the shortcomings of traditional deep learning techniques in such contexts [29].

Data augmentation is a pivotal strategy to mitigate data scarcity, enhancing training example diversity without new data collection. Techniques like back-translation and noise injection generate synthetic corpora, enriching training datasets for models in low-resource environments [15, 28].

Unsupervised alignment, crucial for languages without whitespace-delimited orthography and minimal gold-standard data, facilitates data alignment, such as sentence alignment, without labeled examples, thereby aiding low-resource language processing [17].

Language model adaptation involves fine-tuning pre-trained models for specific tasks or domains, often challenged by the scarcity of large datasets needed for low-resource languages [13]. Effective adaptation is vital for enhancing model performance across diverse linguistic contexts, particularly in languages with limited annotated data [13].

In multilingual settings, developing systems capable of performing NLP tasks across various languages while managing a constrained annotation budget is a significant challenge. Benchmarks assessing multilingual instruction-following capabilities, especially those focusing on multi-language translation, require semantic alignment and effective in-context learning by large language models (LLMs) [28]. Additionally, multitask learning, which involves modeling multiple language generation tasks simultaneously, improves efficiency and performance in low-resource NLP applications [24].

These methodologies and definitions underscore the complexities and innovative strategies in low-resource NLP, addressing data scarcity and linguistic diversity. The field continues to evolve, integrating advanced techniques to overcome the limitations inherent in low-resource settings [25].

In recent years, the fields of few-shot and zero-shot learning have garnered significant attention due to their potential to address data scarcity challenges, particularly in low-resource natural language processing (NLP) environments. To elucidate the intricate relationships among these concepts, Figure 2 illustrates the hierarchical structure of key concepts in few-shot and zero-shot learning. This figure highlights the roles of meta-learning, task augmentation, and zero-shot learning approaches, thereby providing a comprehensive overview of how these methodologies interact and contribute to

overcoming the limitations posed by limited data availability. Such a visual representation not only enhances our understanding of these complex interrelations but also serves as a valuable reference for researchers aiming to explore innovative solutions in the realm of NLP.

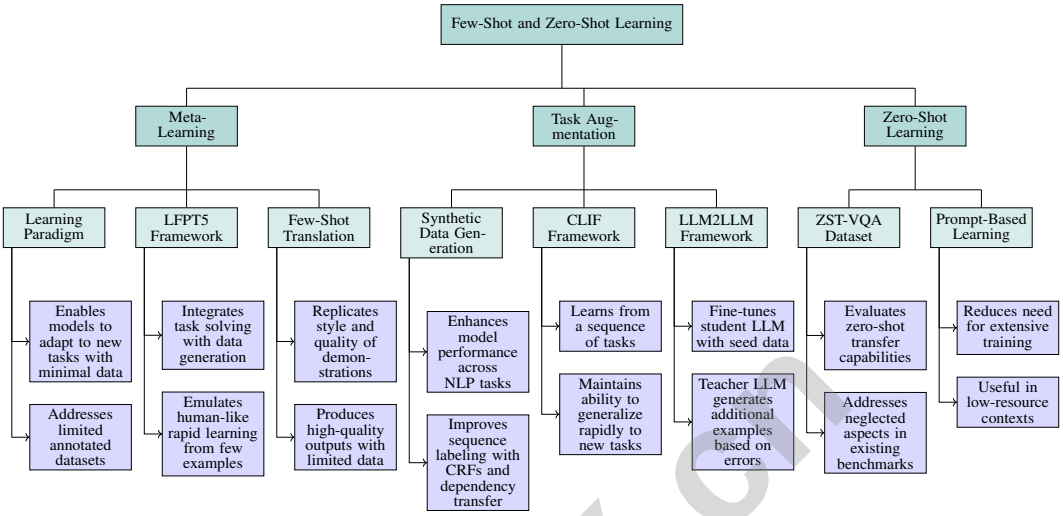


Figure 2: This figure illustrates the hierarchical structure of key concepts in few-shot and zero-shot learning, highlighting the roles of meta-learning, task augmentation, and zero-shot learning approaches in addressing data scarcity challenges in low-resource NLP environments.

3 Few-Shot and Zero-Shot Learning

3.1 Meta-Learning and Task Augmentation

Meta-learning and task augmentation are crucial for advancing few-shot and zero-shot learning in low-resource NLP. Meta-learning, through its "learning to learn" paradigm, enables models to adapt swiftly to new tasks with minimal data, effectively addressing the challenges of limited annotated datasets. The LFPT5 framework exemplifies this by integrating task solving with data generation, emulating human-like rapid learning from few examples [30]. This approach is particularly beneficial in few-shot translation, allowing models to replicate the style and quality of given demonstrations, producing high-quality outputs despite limited training data [5].

Task augmentation enhances model performance by generating synthetic data across various NLP tasks. The integration of conditional random fields (CRFs) with dependency transfer mechanisms significantly improves sequence labeling under few-shot conditions [31]. The Continual Learning of Few-Shot Learners (CLIF) framework enables models to learn from a sequence of tasks while maintaining the ability to generalize rapidly to new tasks [32].

In zero-shot learning, the ZST-VQA dataset introduces a novel approach by structuring training and testing sets to evaluate zero-shot transfer capabilities, an often neglected aspect in existing benchmarks [33]. Prompt-based learning methods further reduce the need for extensive training, proving especially useful in low-resource contexts [6].

The LLM2LLM framework demonstrates the efficacy of task augmentation by fine-tuning a student large language model (LLM) and assessing its performance on seed data, with a teacher LLM generating additional examples based on identified errors [10]. These methodologies collectively address data scarcity challenges in low-resource NLP, improving the applicability and performance of models across diverse linguistic environments.

As illustrated in Figure 3, this figure classifies key methodologies in Meta-Learning and Task Augmentation within NLP, focusing on Meta-Learning frameworks like LFPT5, Task Augmentation techniques such as CRFs with dependency transfer, and Zero-Shot Learning datasets like ZST-VQA.

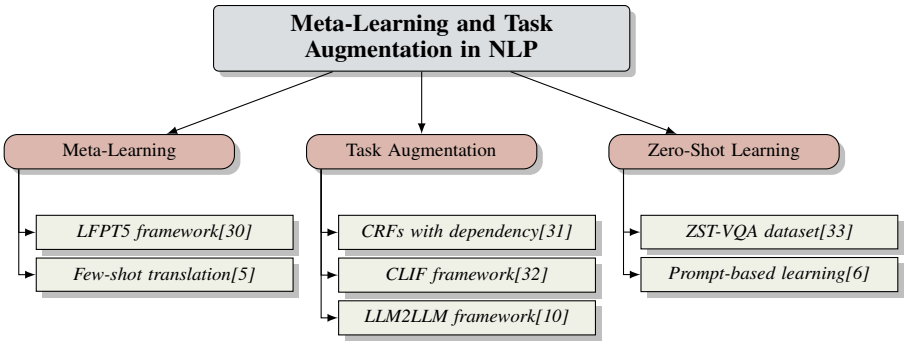


Figure 3: This figure illustrates the classification of key methodologies in Meta-Learning and Task Augmentation within NLP, focusing on Meta-Learning frameworks like LFPT5, Task Augmentation techniques such as CRFs with dependency transfer, and Zero-Shot Learning datasets like ZST-VQA.

4 Data Augmentation Techniques

Data augmentation is crucial for addressing the challenges of low-resource natural language processing (NLP) by increasing the volume, quality, and diversity of training data, thereby enhancing model training. Techniques like answer span manipulation and text simplification generate more representative samples. Advanced methods such as the Knowledge Mixture Data Augmentation Model (KnowDA) leverage task-specific knowledge to produce high-quality synthetic data, boosting pre-trained language models' performance across various NLP benchmarks [34, 35, 36, 37]. Synthetic data generation and self-training are pivotal in augmenting training datasets in low-resource contexts, significantly improving model performance.

4.1 Synthetic Data Generation and Self-Training

Synthetic data generation and self-training are essential for enhancing model performance in low-resource NLP settings by increasing training data variability and robustness. Techniques such as back-translation effectively generate synthetic parallel data for low-resource languages [15]. The Chain-of-Dictionary Prompting (COD) framework augments translation prompts with multilingual dictionary chains to facilitate contextual translations [25].

The LFPT5 framework uses pseudo samples from previous tasks to aid knowledge retention and reduce forgetting, crucial for continuous learning in low-resource settings [30]. Few-shot translation using a transformer decoder-only model trained with self-supervised learning demonstrates effectiveness in overcoming data scarcity [5].

Self-training, which uses model predictions as pseudo-labels for unlabeled data, is exemplified by the LLM2LLM framework, which iteratively augments a small seed dataset with synthetic examples from incorrect predictions by a student large language model (LLM) [10]. This process enhances datasets and model performance in low-resource scenarios.

Integrating token similarities and label dependencies in few-shot sequence labeling tasks, as shown by [31], highlights synthetic data generation's potential. The continual learning framework proposed by [32] supports few-shot adaptation by generating adapter weights for a frozen BART model using bi-level task representations, facilitating continual learning and minimizing forgetting.

These methodologies advance low-resource NLP by addressing data scarcity. Techniques like the "generate, annotate, and learn" (GAL) framework and back-translation significantly enhance model performance in low-resource environments. GAL uses language models to create high-quality synthetic text, annotated with pseudo labels for effective knowledge distillation and self-training, achieving state-of-the-art results in various NLP tasks. Tigrinya translation research shows back-translation through a higher-resource language effectively generates auxiliary data, improving neural machine translation performance in low-resource scenarios [15, 38].

4.2 Continuous Semantic and Text-Based Augmentation

Continuous semantic and text-based augmentation techniques enrich data variability, especially in low-resource NLP scenarios. The KnowDA framework exemplifies this by generating diverse synthetic data through a multi-task training approach and an auto-regressive generation framework, enhancing training datasets' diversity and contextual relevance [35]. This is particularly beneficial where annotated data is scarce, allowing models to learn from a richer set of examples.

The LLM2LLM framework demonstrates targeted augmentation's effectiveness by focusing on challenging examples, leading to more efficient learning and improved performance in low-data scenarios [10]. This approach ensures augmented data is diverse and strategically enhances the model's ability to handle difficult cases, improving overall robustness.

Incorporating pseudo samples into training, as seen in the LFPT5 framework, involves generating synthetic data from previously learned tasks and combining it with new task data. This method facilitates continuous learning and adaptation by tuning prompt embeddings for new task types, enhancing model generalization across tasks and domains [30].

Benchmark protocols introduced by [39] emphasize using few labeled examples alongside unsupervised criteria for model selection, highlighting minimal supervision's potential for effective model performance in low-resource settings.

These continuous semantic and text-based augmentation techniques collectively advance low-resource NLP by enhancing data variability and improving model robustness. By employing advanced methodologies to generate diverse and contextually relevant synthetic data, researchers effectively address NLP data scarcity. Approaches like the "generate, annotate, and learn" (GAL) framework facilitate high-quality task-specific text creation through fine-tuning language models and innovative techniques like knowledge distillation and self-training. Studies indicate synthetic data can serve as a reliable benchmark for simpler tasks, but its effectiveness varies for complex applications, emphasizing the need for data from multiple larger models to mitigate biases. These methodologies enhance NLP models' learning and application across various linguistic contexts, particularly in specialized or low-resource domains [40, 36, 38].

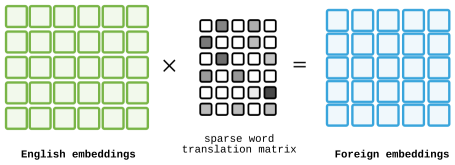
5 Cross-Lingual and Multilingual Transfer

5.1 Techniques for Cross-Lingual Transfer

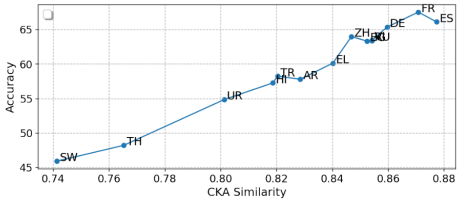
Cross-lingual transfer techniques are pivotal for extending NLP capabilities to low-resource languages by utilizing high-resource language data. The Multilingual Neural Machine Translation (MNMT) model leverages linguistic similarities, such as those among Ge'ez-related languages, to improve translation quality in low-resource scenarios [1]. An empirical evaluation of multilingual pre-trained language models (mPLMs), including mBART and NLLB-200, underscores their efficacy in cross-lingual transfer across languages [41]. The introduction of translation pair prediction (TPP) as a pretraining task enhances mBERT's zero-shot multilingual transfer capabilities [42].

Despite these advancements, achieving language-neutral representations remains challenging. mBERT's limitations in semantic transfer tasks highlight the need for improved language neutrality [43]. Few-shot learning has been shown to improve translation quality with minimal resources, making it particularly relevant for cross-lingual applications [5]. By leveraging bilingual data and shared linguistic features, these methodologies enhance transfer effectiveness. Factors such as subword overlap and word order significantly influence zero-shot transfer performance, with strong correlations observed between transfer success and word embedding alignment. While massively multilingual pretrained models have shown effectiveness in zero-shot transfer, many low-resource languages remain underrepresented. Novel approaches are being developed to identify languages that do not benefit from cross-lingual models, assessing model sensitivity using unlabeled text [44, 45].

As illustrated in Figure 4, methodologies for cross-lingual transfer are exemplified by the "Sparse Word Translation Matrix," which visualizes the structured alignment of words across languages, and the "Accuracy vs. CKA Similarity" plot, which demonstrates complexities in effective cross-lingual and multilingual transfer [46, 47].



(a) Sparse Word Translation Matrix[46]



(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.[47]

Figure 4: Examples of Techniques for Cross-Lingual Transfer

5.2 Multilingual Pretrained Models

Multilingual pretrained models (MMPLMs) advance cross-lingual transfer by leveraging linguistic knowledge across diverse languages. Models like mT5, mBART, and NLLB-200 excel in zero-shot transfer tasks by capturing syntactic and semantic nuances. mBERT's encoding of cross-linguistic syntactic differences impacts its zero-shot transfer performance [48]. Frameworks utilizing transferable phoneme embeddings create a unified latent space for phonemes across languages, enhancing cross-lingual adaptation [49]. The SMALA method improves alignment and performance in cross-lingual tasks by addressing challenges in bilingual subword vocabularies [50]. A centering procedure enhances language neutrality within mBERT representations, crucial for cross-lingual transfer [43]. A benchmark analysis indicates that supervised approaches outperform in-context learning in multilingual natural language understanding tasks, emphasizing refined pretraining strategies [9].

MMPLMs like mBERT and XLM-R address linguistic diversity challenges, facilitating zero-shot transfer without extensive retraining. Recent developments focus on larger models, comprehensive benchmarks, and techniques to enhance performance on unseen languages. Empirical studies demonstrate competitive results in cross-lingual tasks, emphasizing the importance of factors like word embedding alignment in improving transfer performance. By leveraging shared linguistic features and enhancing language neutrality, these models expand NLP applicability across diverse languages and tasks [51, 46, 3, 41, 45].

5.3 Role of Initiatives like NLLB

Initiatives like No Language Left Behind (NLLB) are crucial for advancing cross-lingual transfer by addressing low-resource language challenges and promoting digital inclusivity. NLLB's development of comprehensive multilingual datasets, such as MultiEURLEX, facilitates cross-lingual classification research [52]. This initiative demonstrates substantial improvements in translating low-resource languages [19]. The performance of advanced multilingual models like NLLB-200 underscores NLLB's contributions to enhancing language technology accessibility [13]. Although challenges remain, such as GPT-4's superior performance in certain translation directions, NLLB remains a critical benchmark for evaluating multilingual machine translation capabilities [28]. The Chain-of-Dictionary Prompting (COD) framework has outperformed the NLLB 3.3B translator in many cases, emphasizing the need for ongoing refinement of multilingual benchmarks and resources to support effective cross-lingual transfer [25]. Future research should integrate additional linguistic features and refine difficulty measurement processes to further enhance cross-lingual transfer and code-switching data generation [48]. These initiatives collectively advance cross-lingual NLP, promoting inclusive and scalable language technologies, improving NLP model applicability across diverse linguistic landscapes.

AI-generated, for reference only.

6 Language Model Adaptation

6.1 Challenges in Adapting Pre-trained Models

Adapting pre-trained models for low-resource languages involves overcoming challenges such as limited annotated data and linguistic diversity. A significant issue is overfitting to the source language during fine-tuning, which limits generalizability [41]. This is compounded by the inconsistent performance of Multilingual Language Models (MLLMs) across languages with scarce resources [3]. Few-shot learning methods often fail to adequately address the need for simultaneous classification of new and fixed classes [53]. Language-neutral components in models like mBERT may lack robustness in tasks such as machine translation quality estimation [43]. Pre-trained models do not consistently outperform meta-learning models, with dataset diversity heavily influencing outcomes [34].

The adaptation process is further hindered by the reliance on large development sets for early stopping, leading to inflated performance estimates in low-resource NLP [4]. Limited task-specific data exacerbates overfitting, resulting in the loss of previously acquired knowledge [30]. Catastrophic forgetting remains a challenge, as models degrade in performance on earlier tasks when new tasks are introduced [32]. Data contamination and limited availability of shots in few-shot learning further restrict model efficacy [6]. These challenges highlight the need for innovative strategies to enhance pre-trained model adaptability, focusing on bias reduction, transparency in model training, and robust data augmentation techniques.

6.2 Techniques for Rapid Adaptation

Rapid adaptation of language models to new tasks and domains is essential in low-resource NLP, where data scarcity is a major obstacle. Techniques such as the LFPT5 framework enhance adaptability by retaining knowledge through pseudo sample generation and minimizing forgetting via prompt embedding tuning, thus supporting continuous learning [30]. The LLM2LLM framework exemplifies rapid adaptation by improving large language model performance in low-data scenarios through iterative augmentation of small seed datasets with synthetic examples [10]. Unified Prompt Tuning (UPT) enhances generalization by learning from diverse task groups, crucial in low-resource settings. Subword Mapping and Anchoring across Languages (SMALA) addresses cross-lingual adaptation by leveraging subword similarities to construct bilingual vocabularies, improving multilingual model performance in tasks like cross-lingual natural language inference (XNLI) and neural machine translation [54, 50].

Future research should develop sophisticated unsupervised methods addressing morphological complexity and leverage weak supervision to enhance rapid adaptation techniques. Alternatives to extensive development sets for early stopping could improve adaptation efficiency, as traditional sets are impractical in low-resource contexts. Research indicates reliance on these sets can lead to performance estimation inaccuracies, with discrepancies up to 18

These methodologies advance low-resource NLP by providing robust solutions for rapid model adaptation. By aligning embeddings, employing synthetic data generation frameworks like "generate, annotate, and learn" (GAL), and implementing data augmentation and ensemble techniques, these strategies enhance model adaptability and performance across diverse tasks and domains, particularly in addressing data scarcity in specialized fields [40, 36, 38].

6.3 Knowledge Distillation and Model Optimization

Knowledge distillation and model optimization are vital for enhancing language model adaptability in low-resource NLP tasks, focusing on transferring knowledge from larger models to smaller, efficient ones to improve performance while reducing computational demands. The LetzTranslate approach exemplifies this by producing high-performance bilingual translation models with lower computational requirements, beneficial in low-resource settings [55]. Knowledge distillation uses a teacher model to guide a smaller student model in learning intricate representations, enhancing understanding of decision-making processes. Methods like TeacherLM annotate fundamental concepts and common errors to improve comprehension. In-context learning distillation combines language modeling objectives with in-context examples, enabling smaller models to acquire few-shot learning capabilities from larger models while reducing data demands typical of deep learning [56, 57, 58, 23, 29]. This

technique facilitates knowledge compression without significant performance loss, making it suitable for low-resource environments.

Model optimization refines internal representations to better align with target languages. The LMS method ranks models based on target language performance, enhancing cross-lingual transfer beyond traditional English validation methods [51]. Techniques like soft layer selection in meta-learning frameworks reduce the need for extensive meta-parameters, showcasing efficiency compared to approaches like X-MAML [59]. This reduction optimizes model architecture for specific tasks in low-resource scenarios.

These strategies enhance low-resource NLP by providing effective solutions to model adaptation challenges, emphasizing data augmentation techniques that prioritize quality over quantity [12, 26]. Through strategic application of knowledge distillation and model optimization, these methodologies improve the efficiency and performance of language models, facilitating their application across diverse linguistic landscapes with limited resources.

7 Conclusion

7.1 Innovations and Future Directions

Advancements in low-resource NLP are essential to overcoming the challenges posed by limited data availability and linguistic diversity. Future research should focus on refining inference-time control methods to enhance model adaptability across diverse linguistic contexts. Additionally, the development of robust techniques for generating low-resource datasets and broadening the scope of NLP tasks is crucial for advancing the field and mitigating methodological biases. The LFPT5 framework offers a promising path by improving the generation of pseudo samples and evaluating task adaptability, thus strengthening the resilience of continual learning models. Furthermore, efforts to reduce model size while leveraging multilingual capabilities in few-shot learning could facilitate the integration of larger datasets without sacrificing the benefits of minimal data requirements.

In sequence labeling, enhancing dependency transfer mechanisms and exploring additional contextual embeddings can improve model performance in complex scenarios. Moreover, refining continual learning algorithms to integrate seamlessly with large-scale pre-trained models and investigating task-agnostic scenarios may lead to more flexible and efficient learning frameworks. Future research could also explore additional dialogue tasks, improve prompt engineering, and incorporate human assessments in benchmark evaluations to enhance the applicability of few-shot learning in conversational AI. By pursuing these innovative strategies, future research can effectively address the current limitations in low-resource NLP, thereby expanding the reach and impact of NLP technologies across underrepresented 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] 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.

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

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

[31] 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.

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

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

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

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

[37] 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.

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

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

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

[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] Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. When is bert multilingual? isolating crucial ingredients for cross-lingual transfer, 2022.

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

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

[48] 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.

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

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

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

[52] 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.

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

[54] Pavel Efimov, Leonid Boytsov, Elena Arslanova, and Pavel Braslavski. The impact of cross-lingual adjustment of contextual word representations on zero-shot transfer, 2023.

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

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

[57] Tasmia Shahriar, Kelly Ramos, and Noboru Matsuda. Assertion enhanced few-shot learning: Instructive technique for large language models to generate educational explanations, 2024.

[58] David Isele, Mohammad Rostami, and Eric Eaton. Using task descriptions in lifelong machine learning for improved performance and zero-shot transfer, 2017.

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

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.