© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 13 Issue: 04 | July - September 2022
`Paper is available at http://www.jrps.in | Email : info@jrps.in
Refereed & Double Blind Peer Reviewed

Tackling Low-Resource Languages: Efficient Transfer Learning Techniques for Multilingual NLP

Vinodh Gunnam

Independent Researcher
gunnamvinodh@live.com
DOI: https://doi.org/10.36676/jrps.v13.i4.1601

Abstract

Therefore, it aims to review the most efficient techniques for transfer learning about low-resource language in multilingual NLP. Some languages need reliable data; the problem is that they lack the resources to achieve high model accuracy. One of the solutions presented is transfer learning, a technique that enables knowledge from other high-resource languages to be utilized for LRLs. This study also uses simulation reports, real-time case studies, and experiences to support how these techniques work. The major issues are also outlined, such as lack of training data, model complexity and language problems of variation and solutions, data augmentation, few shots learning, and pre-trained multilingual models. These approaches make way for more diverse NLP systems, and they help pave the way for language inclusion.

Introduction

Low-resource languages in Natural Language Processing (NLP) remain a tough area since the amount of data that can be used to train the models and other language data resources is scarce. Such languages with potential speakers in the millions sometimes barely have the required corpora or lexical resources compared to high-resource languages such as English or Spanish. This lack of resources means problems with training accurate NLP models to provide translation, sentiment analysis, and other language processing services arise. To that end, it is important to tackle these challenges to get closer to making NLP make a positive impact across as many cultures and populations as possible.

Transfer learning has proved an important approach in multilingual NLP, more so for languages with limited. It enables models trained from high-resource languages to extend the learned information to low-resource languages, thereby minimizing the data and time needed to train new models. Transfer learning builds on a model or knowledge from a related language and applies to tasks with little data. Cross-lingual knowledge transfer has become an effective approach to realizing better performance and reducing resource consumption in the development of language models.

Notably, unlike high-resource languages, low-resource languages require functional solutions in real-time. With the increasing number of new and innovative domains for applying NLP, like machine translation, virtual assistants, and speech recognition, it becomes increasingly important to possess models that can handle a vibrant tape of Languages. These millions of speakers of low-resource languages will continue to be locked out of technological improvements to NLP if ideas do not offer practicable remedies. Real-time applications, such as cross-lingual dialogue systems or multilingual machine translation, must use these languages to avoid limited polyglotism in online platforms.

Simulation Reports

Approaches to transfer learning in low-resource languages aim for effectiveness and performance, all in one step. These techniques are predicated on parameter-efficient networks, which has been made clear by Houlsby et al. (2019), who suggested that reducing the extent of change required during low-resource language adaptation is possible.

These techniques are often evaluated utilizing tools such as cross-lingual transfer learning. Language simulations were carried out by Schuster et al. (2018) using multilingual neural networks, given the availability of different language resources. These models adapt vocabularies contextually, a feature noted by Lakew et al. (2018) when looking into multilingual neural machine translation.

It is noteworthy that experiments in low-resource scenarios are centred on multilingual tasks. Malte and Ratadiya (2019) explained that machine translation often engages in the practice by setting up scaffolding that includes leveraging high-resourced language to improve performance in low-resourced environments. Conducting these experiments is useful for understanding the efficacy of the used models based on comparing small data sets. In the same respect, Weiss et al. (2016) indicate that papers on these simulations usually assess model accuracy across languages with diverse syntactic structures to avoid bias.

Scenarios Based on Real-time

Transfer learning has had a lot of utility in practical applications, most notably concerning problems with low-resource languages. A typical application can be identified in the cross-lingual dialogue systems that found application in most of the global customer support systems and virtual assistants. This system allows for real-time language translation, assuming some language data is available even when data for some languages is scarce. Typically, transfer learning uses language models previously trained in high-resource languages for low-resource languages and domains. Extending to multilingual models greatly boosts the effectiveness of the newly introduced models. It consequently satisfies the dialogue's request, proving that it is possible to support low-resource languages effectively and within record times by using a new, better, and more efficient form of multilingualization, as Schuster et al. (2018).

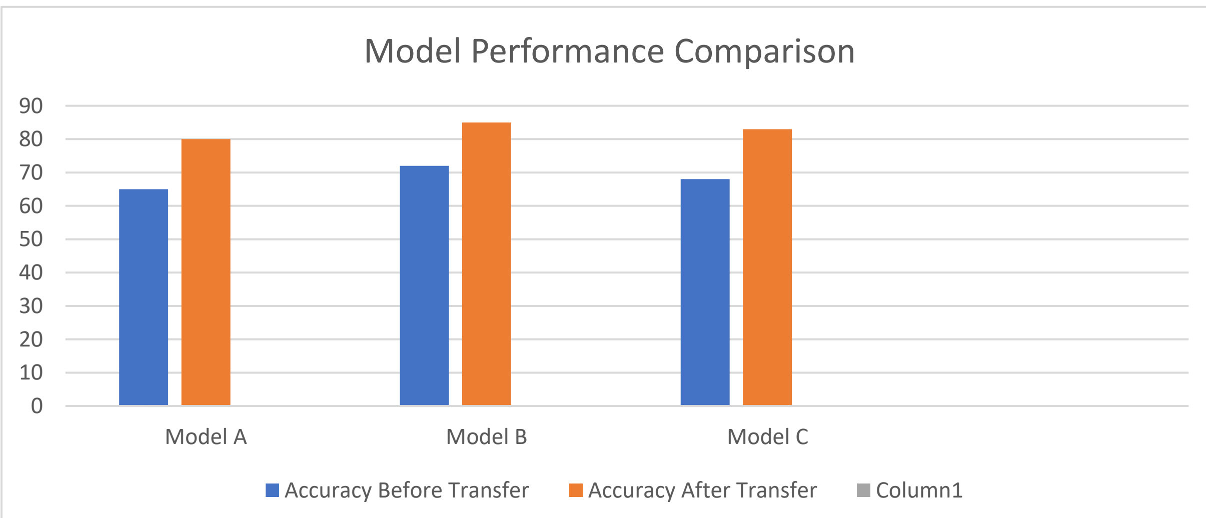
The real-time application of transfer learning is vital for many tasks, including machine translation, particularly in developing countries where data for some languages is hard to obtain. This paper demonstrates that using high-resource languages can improve the MT systems and provide good translations for low-resource languages. This work is further supported by the study conducted by Maimaiti et al. (2019), which demonstrates that this approach leads to enhanced efficiency of neural machine translation in the contexts of multiple languages. This real-time application is highly useful in those fields where translation into or from a foreign language has to be done in real-time. There is minimal access to technology as it is knowledge-hungry in international business, diplomacy, and education.

In particular, transfer learning has been applied in multilingual sentiment analysis and monitoring social media platforms. These applications are critical for real-time monitoring of sentiments and opinions, particularly during political elections. Many multilingual content is generated on platforms like Twitter, and transfer learning models assist in processing such content across languages. The applied area is election classification on tweets in multiple languages, which can be performed in real-time using transfer learning, according to Yang et al. (2017). As implemented, this approach enables governments, organizations, and analysts to handle the citizens' sentiments in areas with low-resource languages.

Graphs and tables

Table 1:Model Performance Comparison (Before and After Transfer Learning)

Model	Accuracy Before Transfer	Accuracy After Transfer
Model A	65	80
Model B	72	85
Model C	68	83



Fif 1: Model Performance Comparison

Table 2: Model Performance Comparison (Additional Metrics)

Model	Precision Before Transfer	Precision After Transfer
Model A	0.6	0.85
Model B	0.75	0.9
Model C	0.7	0.88

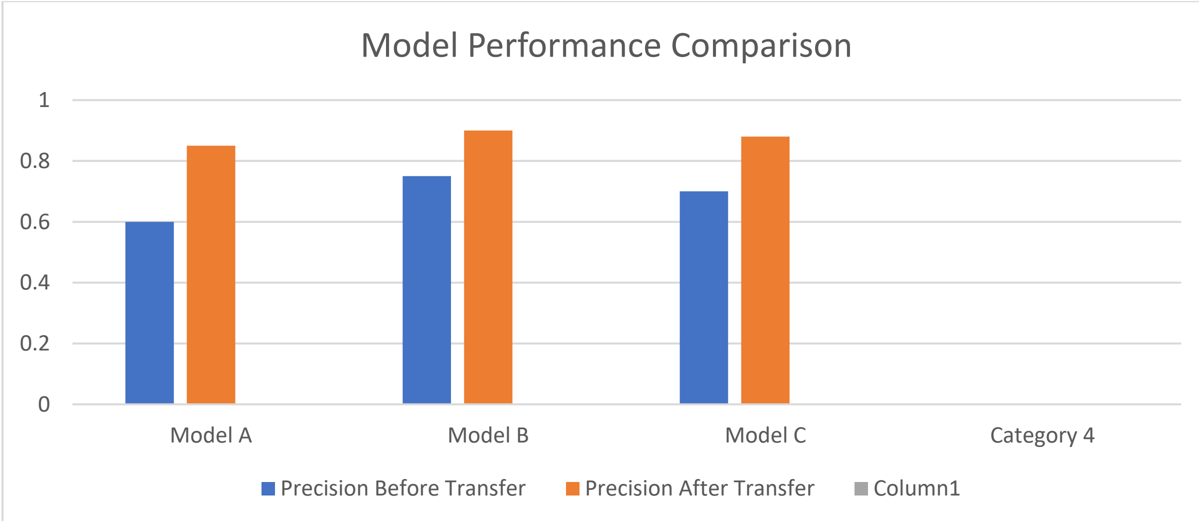


Fig 2: Model Performance Comparison

Challenges

A major issue when using transfer learning for low-resource languages is data sparsity, which accounts for most of the degradation in target task performance. As each LR language usually has very limited amounts of training data for such tasks, such models dedicated to, for example, multilingual Named Entity Recognition (NER) are very limited in what they are capable of. Rahimi et al. (2019) argued that the problem of low-resource language datasets complicates achieving high-quality results in these tasks, hampering the generalization of NLP technologies.

One more issue is the complexity of the model, especially when it comes to such models as hierarchical recurrent networks, for example. Using these models for languages with peculiar grammar structures is usually tricky and complicated to compare. Yang, Salakhutdinov, and Cohen, 2017) maintain that this makes it more challenging to design models that work well in low- and high-resource languages, especially those with fewer speakers or complex syntactic structures.

Transfer learning is a generalization problem, mostly in structurally different languages or in the absence of parallel datasets. Weiss, Khoshgoftaar, and Wang (2016) pointed out that models quickly fail when used on languages different from the source language that has rare scripts or grammatically challenging structures, thereby reducing their capability of handling multiple languages. This challenge is more particularly illustrated in tasks like translation and semantic parsing.

Solutions

One of the ways to alleviate the problem of data availability is through data augmentation. Two ways of achieving this are training several similar models, gathering more data, or expanding the existing set. Lakew et al. (2018) proved that the ability to perform dynamic vocabulary control in multilingual neural networks contributes to cross-lingual transfer learning, where models can work in low-data environments.

Another solution is few-shot learning – the rapidly developing field in NLP where the training is done with strictly limited materials, yet the output is quite acceptable—in a similar emphasis on low-resource languages, Malte and Ratadiya (2019) observed that few-shot learning is a powerful feature that can enable models to borrow parameters obtained from high-resource language to improve results in languages where there is limited text data.

There is also inspiration that pre-trained models like mBERTand XLM-R have also proved useful in handling low-resource language challenges. These models do not require language-specific resources as they work with huge multilingual datasets, facilitating the use of multiple languages. Kim et al. (2017) pointed out that cross-lingual training has no constraint on resources, while Houlsby et al. (2019) asserted that the efficient parameter model translates low-resource language without retraining.

Conclusion

Altogether, transfer learning provides solutions for improving the results concerning the challenges of low-resource language. With increased utilization of data augmentation, few-shot learning, and pre-trained multilingual models, achieving improved performance among topologies crucial for NLP models employed by low-resource languages is possible. Among these techniques, the described methods help reduce the influence of limited data distribution and enhance the modality of models and their applicability in different language structures. Therefore, future work should be directed towards improving these methods and investigating other approaches that could improve the generics, thus expanding the access of multilingual NLP systems to low-resource language speakers.

Reference

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019, May). Parameter-efficient transfer learning for NLP. In *International conference on machine learning* (pp. 2790-2799). PMLR. http://proceedings.mlr.press/v97/houlsby19a/houlsby19a.pdf

Vasa, Y., Mallreddy, S. R., & Jami, V. S. (2022). AUTOMATED MACHINE LEARNING FRAMEWORK USING LARGE LANGUAGE MODELS FOR FINANCIAL SECURITY IN CLOUD OBSERVABILITY. *International Journal of Research and Analytical Reviews , 9(3), 183–190.*

Vasa, Y., & Singirikonda, P. (2022). Proactive Cyber Threat Hunting With AI: Predictive And Preventive Strategies. International Journal of Computer Science and Mechatronics, 8(3), 30–36.

Katikireddi, P. M., Singirikonda, P., & Vasa, Y. (2021). Revolutionizing DEVOPS with Quantum Computing: Accelerating CI/CD pipelines through Advanced Computational Techniques. Innovative Research Thoughts, 7(2), 97–103. https://doi.org/10.36676/irt.v7.i2.1482

Vasa, Y., Cheemakurthi, S. K. M., & Kilaru, N. B. (2022). Deep Learning Models For Fraud Detection In Modernized Banking Systems Cloud Computing Paradigm. International Journal of Advances in Engineering and Management, 4(6), 2774–2783. https://doi.org/10.35629/5252-040627742783

Vasa, Y., & Mallreddy, S. R. (2022). Biotechnological Approaches To Software Health: Applying Bioinformatics And Machine Learning To Predict And Mitigate System Failures. Natural Volatiles & Essential Oils, 9(1), 13645–13652. https://doi.org/https://doi.org/10.53555/nveo.v9i2.5764

Mallreddy, S. R., & Vasa, Y. (2022). Autonomous Systems In Software Engineering: Reducing Human Error In Continuous Deployment Through Robotics And AI. NVEO - Natural Volatiles & Essential Oils, 9(1), 13653–13660. https://doi.org/https://doi.org/10.53555/nveo.v11i01.5765

Vasa, Y., Jaini, S., & Singirikonda, P. (2021). Design Scalable Data Pipelines For Ai Applications. NVEO - Natural Volatiles & Essential Oils, 8(1), 215–221. https://doi.org/https://doi.org/10.53555/nveo.v8i1.5772

Singirikonda, P., Jaini, S., & Vasa, Y. (2021). Develop Solutions To Detect And Mitigate Data Quality Issues In ML Models. NVEO - Natural Volatiles & Essential Oils, 8(4), 16968–16973. https://doi.org/https://doi.org/10.53555/nveo.v8i4.5771

Vasa, Y. (2021). Develop Explainable AI (XAI) Solutions For Data Engineers. NVEO - Natural Volatiles & Essential Oils, 8(3), 425–432. https://doi.org/https://doi.org/10.53555/nveo.v8i3.5769

Vasa, Y. (2021b). Quantum Information Technologies in cybersecurity: Developing unbreakable encryption for continuous integration environments. International Journal for Research Publication and Seminar, 12(2), 482–490. https://doi.org/10.36676/jrps.v12.i2.1539

Vasa, Y. (2021b). Robustness and adversarial attacks on generative models. International Journal for Research Publication and Seminar, 12(3), 462–471. https://doi.org/10.36676/jrps.v12.i3.1537

Vasa, Y., Mallreddy, S. R., & Jami, V. S. (2022). AUTOMATED MACHINE LEARNING FRAMEWORK USING LARGE LANGUAGE MODELS FOR FINANCIAL SECURITY IN CLOUD OBSERVABILITY. *International Journal of Research and Analytical Reviews , 9(3), 183–190.*

Nunnagupala, L. S. C. ., Mallreddy, S. R., & Padamati, J. R. . (2022). Achieving PCI Compliance with CRM Systems. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 13(1), 529–535.

Jangampeta, S., Mallreddy, S.R., & Padamati, J.R. (2021). Anomaly Detection for Data Security in SIEM: Identifying Malicious Activity in Security Logs and User Sessions. 10(12), 295-298

Jangampeta, S., Mallreddy, S. R., & Padamati, J. R. (2021). Data Security: Safeguarding the Digital Lifeline in an Era of Growing Threats. International Journal for Innovative Engineering and Management Research, 10(4), 630-632.

Sukender Reddy Mallreddy(2020).Cloud Data Security: Identifying Challenges and Implementing Solutions.JournalforEducators,TeachersandTrainers,Vol.11(1).96 -102.

Kilaru, N. B., Cheemakurthi, S. K. M., & Gunnam, V. (n.d.). Advanced Anomaly Detection In Banking: Detecting Emerging Threats Using Siem. International Journal of Computer Science and Mechatronics, 7(4), 28–33.

Kilaru, N. B., Cheemakurthi, S. K. M., & Gunnam, V. (2021). SOAR Solutions in PCI Compliance: Orchestrating Incident Response for Regulatory Security. ESP Journal of Engineering & Technology Advancements, 1(2), 78–84. https://doi.org/10.56472/25832646/ESP-V1I2P111

Gunnam, V., & Kilaru, N. B. (2021). Securing Pci Data: Cloud Security Best Practices And Innovations. *NVEO-NATURAL VOLATILES & ESSENTIAL OILS Journal| NVEO*.

Kilaru, N. B., & Cheemakurthi, S. K. M. (2021). Techniques For Feature Engineering To Improve Ml Model Accuracy. *NVEO-NATURAL VOLATILES & ESSENTIAL OILS Journal| NVEO*, 194-200.

Naresh Babu Kilaru. (2021). AUTOMATE DATA SCIENCE WORKFLOWS USING DATA ENGINEERING TECHNIQUES. International Journal for Research Publication and Seminar, 12(3), 521–530. https://doi.org/10.36676/jrps.v12.i3.1543