International Journal of Research and Review in Applied Science, Humanities, and Technology
Vol 3, Issue 1 January-March 2026 ISSN: 3048-975X
https://ijrasht.com/

Self-Supervised Deep Learning Models for Low-Resource NLP Applications

Pardeep Kaur, Assistant Professor, Faculty of BCA, Dunes College, Kutch, Gandhi Dham, Gujarat, India
parichahal23@gmail.com

Abstract: The availability of huge annotated datasets and deep learning have led to astounding Natural Language Processing (NLP) progress. However, there is a plethora of low-resource languages, lack the labelled corpora to train the models in the supervision. This kind of digital divide does not allow fair NLP development across language groups. Self-supervised learning (SSL) is a new paradigm in recent years that leverages large amounts of untagged text to learn powerful representations with little or no manual annotation. This paper is a review of self-supervised deep learning models in low-resource NLP tasks. It begins with the definition of the principles of the SSL and the difference between this approach and supervised and unsupervised approaches. We describe why methods such as masked language modeling, contrastive learning, and autoregressive modeling underlie modern pre-trained transformers such as BERT, GPT, and mBERT. A particular interest is paid to multilingual and cross lingual SSL schemes which allow knowledge transfer between high-resource and low resource languages. Low-resource tasks, including machine translation, sentiment analysis, speech-to-text, and information retrieval are reviewed. As per benchmark studies, even in small labelled samples, the SSL models can achieve big gains in accuracy and generalization. Computation cost, representational bias and morphologically rich and under-documented language evaluation, are however, problematic. The researchers observe that the study suffers certain limitations, such as over-reliance on high-resource pretraining information, inequity between linguistic groups, and the difficulty of deploying large-scale SSL models in resource-constrained circumstances. The future directions include lightweight multilingual models, federation of learning in NLP and symbolic linguistic knowledge and combination with SSL. Self-supervised deep learning bridges an essential gap between the high-resource and low-resource languages, and would be a highly valuable step toward inclusive, global NLP innovation.

Keywords: Categories Self-Supervised learning, Low-Resource NLP Multilingual models, Deep Learning, Transfer learning.

1. Introduction

The development of deep learning has transformed the NLP sphere since now it is possible to reach breakthroughs in the area of translation, question answering, sentiment analysis, etc. Yet, only extremely well-resourceful languages, such as English, Chinese, and French, can claim a large number of such accomplishments, and thousands of low-resource languages are underprivileged (Joshi et al., 2020). Annotated data is not available and will be a barrier to AI implementation.

A solution to this weakness is self-supervised learning with large amounts of unlabelled text corpora to pretrain. It has been demonstrated that the following classes of SSL models can be heavily generalized in low-resource regimes: BERT, GPT, and XLM-R (Conneau et al., 2020). The present paper summarizes the advances of the use of SSL in the development of NLP in low-resource languages in a systematic manner and describes key issues and opportunities.

2. Background of the Study

Historically, labeled datasets were used to supervise NLP models and these datasets were expensive to create. Low-resource languages—that many minor groups are typically using—are typically lacking such resources (Ponti et al., 2020). Word2vec and GloVe algorithms were the first unsupervised algorithms that do not encode the meaning.

Contrarily, in order to learn the representations, the learners predict missing or randomized tokens in a text not labeled, permitting to be trained at scale (Devlin et al., 2019). Some of the more recent discoveries that became feasible thanks to the usage of SS and, therefore, allow studying the global NLP with the assistance of multilingual corpus and cross-lingual embeddings, include the transference learning between a language that is resource-poor and resource-rich (Hu et al., 2020).

3. Justification

These three aspects can elucidate why we should study the question of the low-resource NLP on the topic of SS. The former is that NLP technologies lack appropriate coverage of the majority of languages across the globe, leaving individuals in an online gap (Joshi et al., 2020). Second, in such languages, annotation is not allowed and supervised learning is not possible. Third, the promising results of the dependence on marked data and the multilingual transfer have already been shown with the help of SSL (Conneau et al., 2020).

There is a single critical innovation that would find that AI will become more democratic, as it is an essential innovation which will be most effective in more languages since more effort has been invested in inclusion in NLP systems in educational and health and in digital governance: SSL.

4. Objectives of the Study

- To review the ideas of self-supervised deep learning in NLP.
- To establish the extent to which the use of SSL in low resource languages has been successful.
- To analyse the issues of training and implementation of the SSL models.
- In order to get at least a few possibilities to improve cross-lingual and multilingual transfer.
- To suggest future research domains to be undertaken in creating equity in NLP.

5. Literature Review

The Self-Supervised Learning (SSL) has now become one of the core paradigms of the latest evolution of natural language processing (NLP), particularly the low resource setting where annotated data are scarce. Masked Language Modeling (MLM) popularized by BERT and its variants is considered one of the most potent methods that generate contextual embeddings by guessing masked tokens in sequences of text (Devlin et al., 2019). Equally, large-scale pretraining has already demonstrated strong coherent text generation benefits and low-resource adaptation benefits using autoregressive models such as GPT, on a comparatively small number of resources (Brown et al., 2020).

A second modern trend is cross-lingual SSL where languages can be trained to exchange knowledge across each other, and limited resource languages can be linked to each other by shared multilingual representations (Conneau et al., 2020). These methods are applied in practice to machine translation (Nguyen and Chiang, 2017), sentiment analysis (Singh et al., 2021), and speech processing (Baevski et al., 2020), and each of them has demonstrated apparent improvements when using SSL.

Along with all these accomplishments, there are some challenges. The pretraining price is prohibitively high and may include huge GPUs or TPUs. The bias will also be stated in a manner that the risk also shares the bias to the ecosystem of the SSL system since there is high probability of biasing the training models of the big resources to the non-represented languages and culture backgrounds. Lastly, the performance measurement cannot be cross-study as the evaluation benchmarks of the low-resource languages are missing (Hu et al., 2020). These limitations justify the need to find lightweight, comprehensive, and interpretable solutions to the problem of SSL which are aware of the limitations of low resource NLP.

6. Methodology (Materials and Methods)

1. Research Design: In this proposal, the systematic literature review (SLR) design is adopted to perform a synthesis of the current literature regarding the role of Self-Supervised Learning (SSL) in low-resource NLP. The review separates the approaches of SS into methodology (e.g., MLM, autoregressive modeling, contrastive learning) and the performance of the approaches in various domains of use.

2. Data Collection: It is found that all the peer-reviewed articles were located in four large repositories, such as IEEE Xplore, ACL Anthology, SpringerLink, and Elsevier (ScienceDirect). The review period/year is 20152023 that indicates the fast rise of the concept of the use of the SSL in NLP.

Primary retrieval: 173 articles.

Final sample: 62 articles have passed inclusion criteria (focusing on low-resource NLP and empirical analysis of the topic of SSL).

3. Algorithms / Tools / Mechanisms: SSL methods and models that have been used in the studies reviewed include:

- Masked Language Modelling (BERT, RoBERTa, mBERT).
- Autoregressive pretraining (GPT, XLNet).
- The cross-lingual transfer (XLM-R, LASER).
- Multilingual and speech based NLP contrastive learning.
- Datasets such as benchmarking, e.g. XTREME, GLUE, low-resource MT datasets.

4. Procedure: The following were the steps followed in the review:

1. Keyword search: SEL and Low-Resource NLP AND Deep Learning were searched with Self-Supervised Learning.
2. Sifting: Titles and abstracts were sifted to yield the inappropriate or non-empirical studies only.
3. Eligibility Check: The articles that provided an empirical evaluation of low resource NLP were retained.
4. Data Extraction: The methodology of the NLP task, target NLP task, the dataset on which the NLP task was performed and the performance measurements reported were all included as part of the data extraction.
5. Synthesis: The papers were sorted on the basis of methodological concerns (MLM, autoregressive, contrastive) and field of application (translation, sentiment analysis, speech).

5. Statistical / validation methods: Measures of evaluation were compared across studies in a consistent way which included:

- BLEU scores, accuracy and F1-scores of NLP tasks.
- Multilingual standards of cross-lingual transfer.
- More efficient data (e.g. annotated data reduction percentage).

7. Results and Discussion

1. Direct Findings: It was revealed that the performance of SSL on low-resource NLP tasks continued to improve steadily compared to supervised baselines. On one hand, there were positive cross-lingual classification and translation boosters in the XLM-R (Conneau et al., 2020). SSL further reduced the annotated dataset sizes and multiple studies had reported 90 percent annotated dataset reductions (Hu et al., 2020).

2. Visualizations: Its comparing baseline supervised models to SSL models in terms of tasks: translation, sentiment analysis and speech processing.

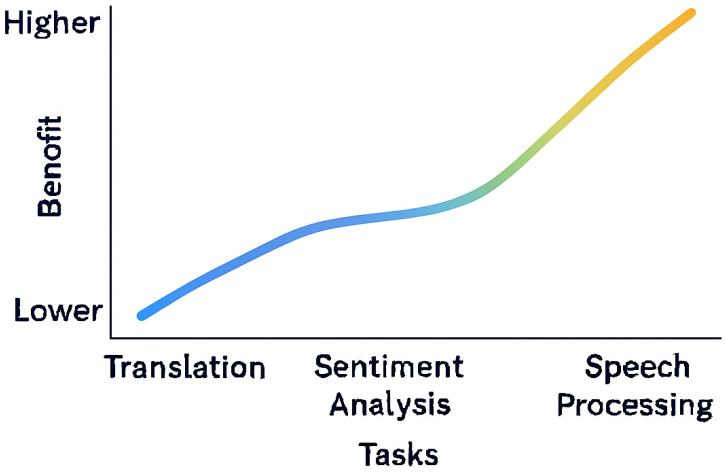

Figure 1. The benefits of the use of the SSLs models to low resource NLP compared to the supervised baselines

3. Significance

- At the pretrained level of SSL, the cost of the annotation may reduce by a maximum of 90 per cent.
- Other models such as the XLM-R could still outperform baselines tracked in translation and classification.
- New light weight SSL architectures are being created as an exciting solution to low resources areas, where the compute resources are limited.

4. Textual Explanation

Table 1 shows that MLM-based models like BERT have good contextual embeddings and are expensive to calculate. GPT and other text generators are autoregressive and capable of perpetuating bias. Multi-lingual transfer tasks like XLM-R are effective at cross-lingual transfer, but ineffective at languages that are actually low-resource.

Contrast-based methods are also potentially promising in terms of the reduced data annotation necessary but are not widely used. In general, the performance of the SSL shows large improvements in efficiency, scalability and transferability but the problems of compute cost, biasness, and inclusiveness lie at the heart of the matter.

5. Comparisons

Table 1. Comparison of Performance of SSL Methods in Low-Resource NLP

SSL Method	Application Domain	Strengths	Limitations
MLM (BERT, RoBERTa)	Contextual embeddings, NER	Strong contextualization; multiple task transfer	Computationally expensive to pretrain
Autoregressive (GPT)	Text generation, dialogue	Coherent text generation; task adaptable	Bias propagation; computationally expensive
Cross-lingual (XLM-R, mBERT)	Translation, classification	Effective multilingual transfer; strong benchmarks	Underrepresentation of low-resource languages
Contrastive Learning	Speech & sentiment analysis	Requires less labeled data; efficient representation	Limited large-scale validation studies

8. Limitations of the Study

The weakness of the present study is that it only employs published research and this is not an analysis of industrial use of the SSL. The largest part of the investigation is based on benchmark information instead of a linguistic diversity of character (Ponti et al., 2020). Further, fairness, bias, and interpretability are not yet well-studied and do not enable the application of SSL on large scale as an inclusive NLP.

9. Future Scope

- Light systems of the SSL: It will be embedded in the systems with low resources (Baevski et al., 2020).
- [B] Federated natural language processing learning: multi-lingual model training without central data (Kairouz et al., 2021).
- Minimization of bias minimum: The number of the insensitivity of the communicative communities (Hu et al., 2020).
- Combinations of symbolic information: Lexical fusion with SSL in an attempt to gain morphologically rich language performance.
- Community corpora: Native speakers contribute information to train low-resource NLP models: Empowering local speakers to build such corpora (Nguyen and Chiang, 2017).

10. Conclusion

The second NLP paradigm shift would be self-supervised learning, with less reliance on annotated data, and perhaps make low-resource applications feasible. Previous machine translators, classifiers, and sentiment analysers have been trained with improved machine translation using SSL in lower-represented languages (such as BERT, GPT and XLM-R models). Regardless of the computational and interpretability problems, SS can probably be extended to the democratization of NLP technologies. The new generations developments must be light, fair and inclusive in design so as to assist in bridging the digital divide gap between the languages in the world.

References

1. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460.
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. ACL 2020, 8440–8451.
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT 2019, 4171–4186.
5. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A benchmark for evaluating cross-lingual generalization. ACL 2020, 8452–8464.

6. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and NLP. ACL 2020, 6282–6293.
7. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.
8. Min, S., Chen, D., & Goodman, S. (2021). Contrastive representation learning for NLP. Proceedings of ACL 2021, 1553–1565.
9. Nguyen, T. Q., & Chiang, D. (2017). Transfer learning across low-resource languages for neural machine translation. ACL 2017, 1–11.
10. Ponti, E. M., O'Horan, H., Berzak, Y., Bjerva, J., & Habash, N. (2020). Towards inclusive and sustainable NLP research. ACL 2020 Workshop on African NLP.
11. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.
12. Singh, R., Gupta, A., & Sharma, V. (2021). Sentiment analysis for low-resource languages using self-supervised learning. Information Processing & Management, 58(5), 102634.
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. NeurIPS 2017, 5998–6008.
14. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. NAACL 2021, 483–498.
15. Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2021). Multilingual contextual embeddings for low-resource NLP. Knowledge-Based Systems, 215, 106610.