iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.61039

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

Advancements in Neural Machine Translation: Methodological Innovations and Empirical Insights for Cross-Linguistic Discourse Preservation

Ethan Lau[1], Reyansh Pallikonda[2], Ethan Alapatt[3], Nathan Aye[4]
First-Fourth Mission San Jose High School

Abstract: ***Natural Language Processing (NLP), in the form recognizable today, really began to take hold in the 1980s, when machine learning helped propel it to soaring heights. However, due to a lack of processing power, machine learning, and to an extent, NLP, started slowing down in innovation and ideas and had almost ground to a relative halt, until the last decade, when a sudden increase in both productivity and interest in the machine learning helped increase the amount of knowledge in the space itself. This review provides several different case studies using different methodologies. The first paper was a deep analysis on how researchers were able to use Tesseract and Google Vision in tandem with automatic data mining methods to enrich the Cherokee language database in order to preserve it from extinction. The second paper takes a query translation-based approach toward translating English to Indian languages and utilizes a Multilingual Cross-Language Information Retrieval (MLCIR) system with tools such as Part of Speech Tagger (POST), Stop-Word, and Porter Stemmer. The third paper presents CoVe, which transfers knowledge from machine translation to improve performance on NLP tasks like sentiment analysis and question answering by using contextualized word vectors along with word embeddings, achieving new state-of-the-art results on some datasets. The fourth paper aims to translate English to Pakistan Sign Language (PSL) and also uses POST and goes through dependency analysis, sentence classification, and PSL using PLS trees. The fifth paper uses a Multilingual Neural Machine Translation (NMT) system for LowResource languages and incorporates two main models: a recurrent NMT and a transformer NMT. The sixth paper analyzes how a fine-tuned transformer model seems to work better than transformer models trained from scratch on high-resource languages, while vice-versa seems to occur for low-resource languages. The seventh paper adds to this by talking about how multilingual translation seems to work better than a back-translation model. Given the diverse array of approaches that could be used, we aim to identify the most efficient and correct methodology for future researchers to use in their work, based on the papers in this literature review.***

Keywords: NLP, Machine Translation, Neural Machine Translation, POS-Tagger, Seq2Seq, LSTM, Bi-LSTM, CNN, RNN

I. METHDOLOGIES

Due to the differing amounts of resources available as well as the phonetic structure of some languages, the translational methodologies employed significantly vary between different languages. For example, the translation methodology for translating between Cherokee and English involves a comprehensive system that leverages various machine learning and natural language processing techniques. The process begins with a machine translation (MT) system that compares parallel text from Cherokee to English, employing automatic data mining methods to gather data. A language identifier and multilingual embeddings support the translation process, which can work with web and OCR (optical character recognition) processed text. The OCR tools, such as Tesseract-OCR and Google Vision OCR API, play a significant role in handling image-based text. The accuracy of OCR tools can be influenced by the quality of images, affecting word and character error rates (WER and CER). Tests with Tesseract and Google Vision show varying levels of accuracy depending on whether the original text or a screenshot is being processed. The system also incorporates automatic speech recognition (ASR), translating audio to Cherokee text using pre-trained and fine-tuned models like XLSR-53. These ASR models handle both audio-to-phonetic text and audio-to-syllabic text with specific word error rates. To achieve accurate tokenization and morphology parsing, the system transliterates text into Latin script for easier learning of morphemes. It uses a collection of gold morphemes from 372 Cherokee words and pretrained tokenizers and parsers to evaluate alignment between subwords and gold morphemes. The methodology also employs unigram language modeling (LM), byte-pair encoding (BPE), and Morfessor, each tested for precision, recall, and F1 score. The methodology also extends to part-of-speech (POS) tagging and dependency parsing to improve the overall quality and context of the translated output.

While this dependency on OCR technology is due to the more picturesque nature of the scripting language, the methodology for more lingual languages differs. For example, AUTHOR details how the methodology used for translating English to Indian language is a Query Translation-based approach inside a Multilingual Cross-Language Information Retrieval (MCLIR) System. The system translates the query-related documents. The system uses a Machine Readable Dictionary called BUBShabdasagar-2011, which serves as a translation lexicon resource available in ISCLL encoding form or plain text (converted to UTF-8/Western Windows encoding). A transliterator is used to overcome out-of-versaab situations using the ITRANS transliteration scheme. A part of speech tagger (POST) classifies words into different parts of speech.

Another differing methodology used was treating the words as vectors and assigning weights to the correlation between the words. AUTHOR describes this methodology for training the machine translation (MT) model for English-to-German translation as a sequence-to-sequence model architecture, specifically utilizing a two-layer, bidirectional Long Short-Term Memory (LSTM) network as the encoder. This model provides contextual information for other natural language processing (NLP) tasks. The procedure begins by feeding word sequences of the source and target languages into the model, with each word corresponding to a GloVe embedding (GloVe(w^x)).

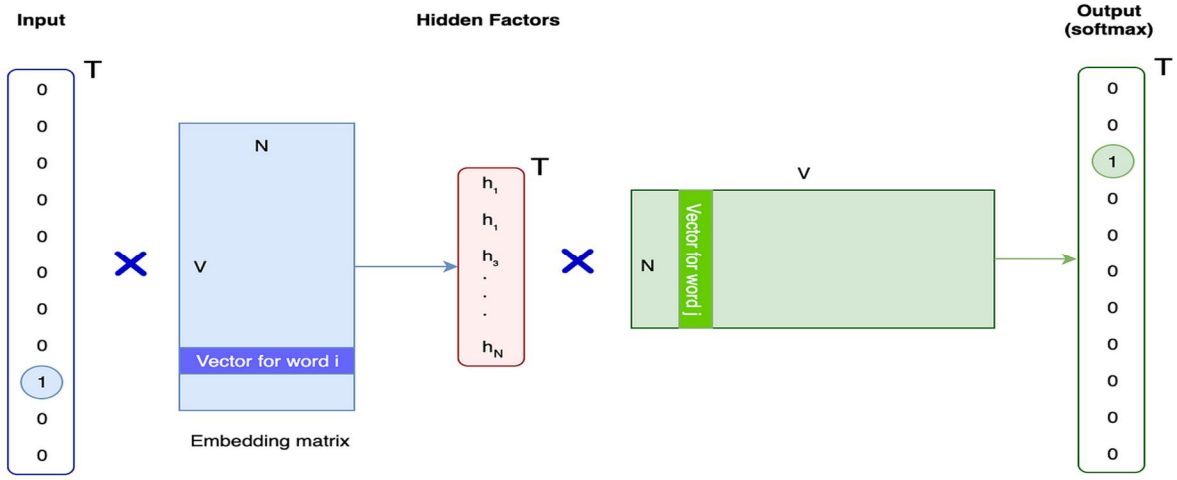


Fig. 1 Glove(w(x)) Embeddings

The decoder then utilizes these embeddings to generate translations, adjusting the hidden states based on context. The tools employed in this process include CoVe (Context Vectors concatenated with Word Vectors from GloVe), which integrates contextual information with word embeddings. The CommonCrawl-840B GloVe model is utilized for English word vectors. Additionally, an attentive classification network is incorporated, employing a feedforward network with Rectified Linear Unit (ReLU) activation and a bidirectional LSTM encoder. A Maxout network is employed to create a probability distribution over possible cases, aiding in decoding and generating translations. Overall, the methodology encapsulates various neural network architectures that create a robust framework to apply for other language techniques. AUTHOR describes a different use of another MT model. They talk about the methodology from translating English to Pakistan Sign Language (PSL) involves the use of a MT Model, specifically a Machine Rule Based Translation Model. This model takes English as input and outputs PSL. The approach refers to the TESSA system. Data used for this model includes collaboration with members of the deaf community and PSl experts to define grammar. Translation is facilitated by a software system that processes a collection of rules using a pre-processing component and dependency analysis utilizing tools such as POS tagging, dependency illustrations, and Stanford parser. The process follows an algorithm with rules that can be updated and modified. PSL tree generation involves manipulation of nodes through deletion, addition, and alteration.

Methodologies for low resource languages like Dravidian languages all utilize Neural Machine Translation for translation tasks as well as the use of advanced model architectures such as LSTM-based NMT and Transformer NMT. These language models all use optimization algorithms such as Adagrad and Adam. These algorithms are used to enhance training performance and efficiency. These methods are specifically used to focus on challenges associated with low-resource languages by leveraging multilingual approaches and back translation. The highlight importance of utilizing contextual information from other languages to improve the model's quality.

The methodology for multilingual neural machine translation (NMT) for low-resource languages focuses on addressing the limitations of traditional NMT systems when translating languages with limited data. This approach utilizes a semantic space across multiple languages and employs a zero-shot self-training method to improve translation quality in low-resource languages. The system encompasses a multilingual NMT model (M) and data (D) for the training process. The two main architectures that are leveraged are LSTM-based NMT and Transformer NMT. An LSTM-based NMT employs Long Short-Term Memory (LSTM) networks for the NMT model, where the architecture includes encoder, decoder, and attention mechanisms trained together. The second architecture, Transformer NMT, utilizes a self-attention mechanism and a stack of encoder-decoder networks that work in an auto-regressive manner, known for its efficiency and superior performance in NMT. The system also integrates optimization algorithms like the Adagrad and Adam to enhance training performance and efficiency. It's designed to handle multiple languages, including English, Italian, and Romanian, with support for six translation directions. Employing a multilingual approach enables the NMT system to effectively translate between languages, leveraging data and contextual information from other languages in the semantic space to improve translation quality, especially for low-resource languages.

The methodologies center on improving Neural Machine Translation (NMT) for low-resource Dravidian languages like Kannada-Malayalam, Kannada-Tamil, Kannada-Telugu, Kannada-Sanskrit, and Kannada-Tulu. They utilize various NMT models such as LSTM, Bidirectional LSTM, Convolutional seq2seq, and Transformer models. Specifically, the methodologies explore the Indic-Indic model for multilingual neural machine translation and CNN trained using the fconv architecture. Results indicate that fine-tuned transformer models with back-translation excel for high-resource language pairs, while transformer models trained from scratch perform optimally for low-resource language pairs like Kannada-Tulu and Kannada-Sanskrit.

Additionally, the methodologies aim to address the challenge of translating low-resource Dravidian languages like Tamil, Malayalam, and Telugu, using dominant NMT paradigms such as encoder-decoder and transformer models. They emphasize the utilization of multilingual translation and back translation approaches to enhance translation performance and robustness. The model architecture, based on the transformer model, consists of six stacked encoder and decoder layers with specific parameters like a layer size of 512, four heads in each attention layer, and a feedforward network size of 1024. Training involves translation in eight directions, employing four GeForce RTX 3090 GPUs with each GPU allocated a batch size of 2,000 tokens. The methodologies aim to improve translation quality for Dravidian languages by addressing the challenges associated with low-resource language translation, providing better performance, and enhancing robustness for these languages.

II. FINDINGS

Based on the multiple experiments reviewed, the respective results and performance vary due to their differing methodologies/techniques, available resources, and the languages utilized.

The English to Cherokee Language Translation experiment tested the Tesseract-OCR and Google Vision OCR API tools for data mining. The “Original” set contains 20 images of unedited pages of children’s books while the “Screenshot” set contains 20 edited images that omit background noise from the photos, effectively improving image quality. (Word Error Rate = WER, Character Error Rate = CER)

OCR Tools	Original (WER)	Original (CER)	Screenshot (WER)	Screenshot (CER)
Tesseract	0.355	0.230	0.151	0.063
Google Vision	0.533	0.199	0.468	0.074

Fig. 2 OCR Tool Comparison

The table indicates that the results of Google Vision generally have a higher WER and CER compared to the Tesseract’s. Additionally, the “Screenshot” set has a lower WER and CER than the “Original” set on average, which is logical as OCR tools heavily depend on image quality. Given these decent results, they can be greatly improved if image quality is prioritized. This paves a path for a possible future of OCR tools as error rates seem inversely proportional to image quality. Another tool used in this experiment that was tested was the XLSR-53, a pre-trained Automatic Speech Recognition (ASR) model with 53 stored languages. Testing the Cherokee ASR models of audio-to-phonetic text and audio-to-syllabic text with provided datasets by Michael Conrad, the experimenters achieved the results below:

	Audio to Phonetic Text	Audio to Syllabic Text
WER	0.64	0.21

Fig. 3 Phonetic Text Performance versus Syllabic Text Performance

These results show the Audio to Syllabic text has a better performance than the Audio to Phonetic Text. Accuracy can be further improved with a larger and more representative training dataset. Diving deeper into the more specific tools, the experiment covers the testing of 2 subword tokenizers (Unigram LM and BPE) and a morphology parser (Morfessor). These tools are tested by how subwords and morphemes are aligned in the Cherokee language by testing a dataset of 372 words; the following results are given:

	Precision	Recall	F1
Unigram LM	16.6	19.6	17.9
BPE	14.4	16.5	15.4
Morfessor	16.6	16.3	16.5

Fig. 4 Precision, Recall, and F1 Rate Comparison

Unigram LM seems to perform better than BPE and Morfessor in terms of subword alignments and morphemes. These pre-built tools appear ineffective for subword tokenizers. This may be due to the scarcity of relevant tools based on a niche language such as Cherokee. Overall, this experiment was innovative in the sense of a language approach that is impractical to learn in today's world and provided useful insight for future potential actions to continue this revival.
The Contextualized Word Vectors experiment utilizes a multitude of varying LSTM models, CoVe, and GloVe. The following test was used to determine the validation performances of GloVe, CoVe, and character n-grams.

Dataset	Random	GloVe	Char	CoVe-S	CoVe-M	CoVe-L
SST-2	84.2	88.4	90.1	89.0	90.9	91.1
SST-5	48.6	53.5	52.2	54.0	54.7	54.5
IMDb	88.4	91.1	91.3	90.6	91.6	91.7
TREC-6	88.9	94.9	94.7	94.7	95.1	95.8
TREC-50	81.9	89.2	89.8	89.6	89.6	90.5
SNLI	82.3	87.7	87.7	87.3	87.5	87.9
SQuAD	65.4	76.0	78.1	76.5	77.1	79.5

Fig. 5 Dataset with Embeddings Comparison

It's apparent CoVe greatly aids the validation performance, with CoVe-L being the most effective due to its larger and more diverse training set. GloVe and Char (character n-gram) are relatively similar in results but noticeably less effective than Context Vectors. CoVe seems to be more effective than GloVe due to their respective training processes and the difference in capabilities of their contextualization of words and the depth of the layers. CoVe's training dataset is of a higher quality than GloVe's and CoVe generally has a better grasp of contexts and semantics compared to GloVe. Consistently having the best results, CoVe proves to be the most successful out of the three possible methods.

The English to Pakistan Sign-Language (PSL) experiment utilizes an MT model and methods such as POS tagging and PSL grammar/tense analysis to convert English to PSL. With a testing dataset of 500 sentences, 476 sentences were translated successfully, yielding an accuracy of 95.2%. The 24 sentences that weren't passed were in the complex/complex compound categories. In this case, Contextualized Vectors can help with these failings. With a current BLEU score of 0.78, we believe the added use of CoVe in this experiment will boost the accuracy by a larger margin.

The study explored multilingual neural machine translation (NMT) models to facilitate translation between English, Italian, and Romanian languages. The researchers employed encoder-decoder architectures, evaluating both long short-term memory (LSTM) and Transformer models. A zero-shot self-training approach was adopted, which leverages semantic sharing across multiple languages to improve translation quality. Furthermore, the training process incorporated an iterative train-infer-train methodology, where the models were trained, used to generate translations, and then trained again on the generated outputs. This train-infer-train stage helped to progressively enhance the multilingual NMT systems' translation capabilities.

III. RESULTS

Architecture	Performance
Transformer	Superior to LSTM across all language pairs
Multilingual	Outperformed separate bilingual baselines
Train-Infer-Train	Significantly boosted baseline multilingual NMT

Fig. 6 Architecture Performance Summary

Transformer models were more effective than LSTM across all language pairs, which makes it the most viable model in this paper. Train infer-train process also substantially improved performance when combined with the transformer model. Single multilingual system was better than multiple bilingual systems, and the Train-infer-train process substantially improved performance.

Language Pair	Best model
High resource	Fine tuned transformer + back translation
Low resource	Transfomer from scratch

Fig. 7 Low and High Resource Model Performance.

Back-translation consistently improved translation quality across all high resource languages somewhat. Transformers trained from scratch best for low-resource pairs.

For the low-resource Dravidian languages, it was found necessary to adopt two methods, both multilingual and backtranslation, in order to combat the lack of large- scale annotated parallel data. Using Beam search decoding and BLEU evaluation, these results were evaluated .

System	Review
Multilingual	Higher than bilingual baselines
Back-translation	Highest scores across all directions

Fig. 8 System Architecture Comparison

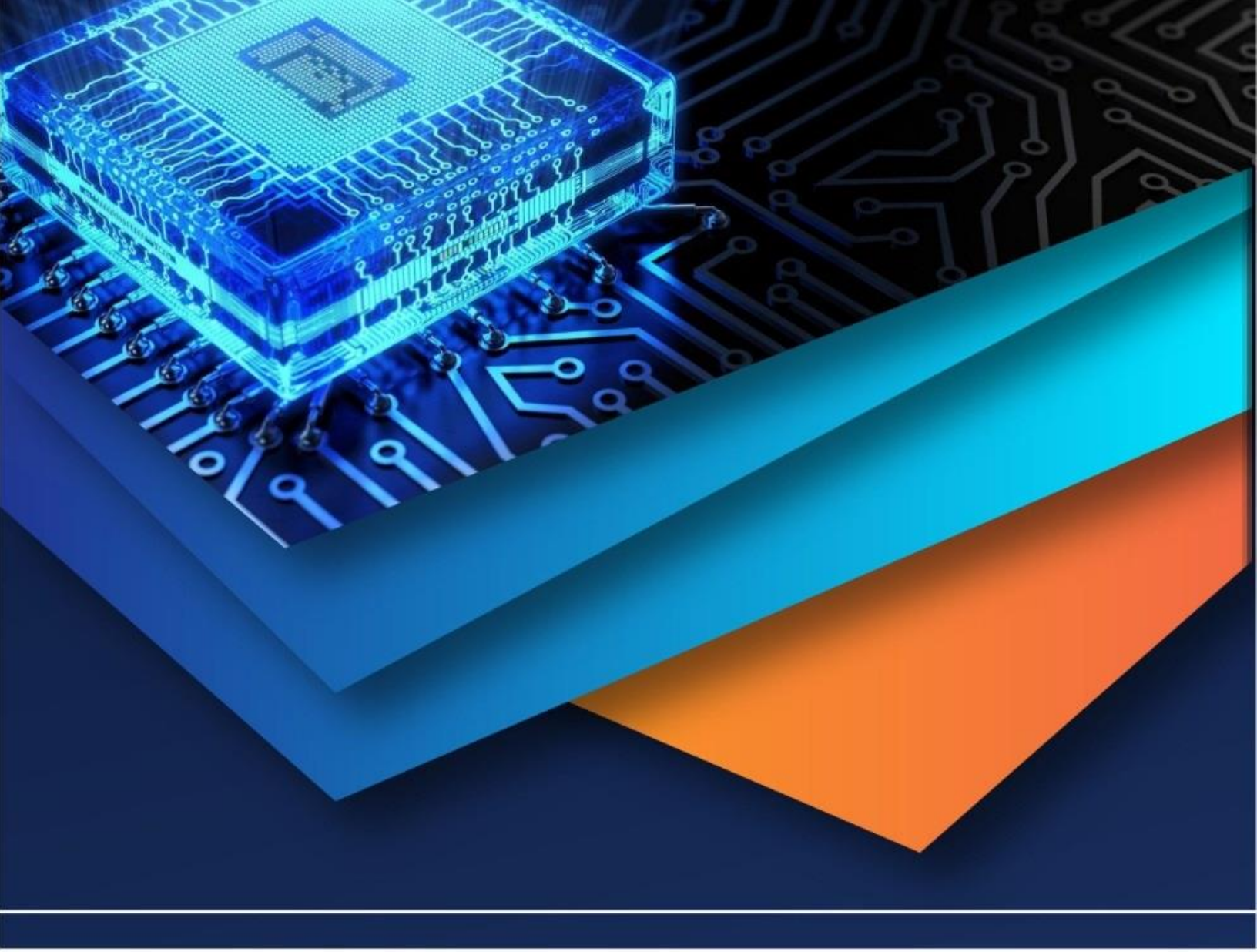
The multilingual Transformer model outperformed bilingual baselines across all 8 Dravidian and English language directions. Incorporating back-translated data into the multilingual model training led to substantial BLEU score improvements, achieving the highest translation quality overall. The back-translation gains were particularly pronounced for lower-resource language pairs with limited parallel data. These results highlight the significant benefits of combining multilingual modeling with back-translation, especially for low-resource machine translation tasks.

IV. CONCLUSIONS

Overall, this study offers a rigorous exploration into the domain of Natural Language Processing (NLP), with a primary focus on machine translation methodologies, notably Neural Machine Translation (NMT). Through a meticulous examination of various case studies, each emblematic of distinct methodological approaches, the research endeavors to distill overarching insights and empirical findings pertinent to the advancement of language processing technologies. The synthesis traverses a trajectory marked by historical antecedents, noting the seminal role of machine learning in the 1980s and its recent resurgence catalyzed by computational advancements and scholarly reinvigoration. Notably, the investigation underscores the transformative potential of deep learning architectures, such as Long Short-Term Memory (LSTM) networks and Transformer models, in enhancing translation fidelity and addressing the exigencies of low-resource linguistic contexts. Across a spectrum of case studies, salient findings emerge, underscoring the efficacy of tailored methodologies in mitigating linguistic constraints. Noteworthy instances include the utilization of Optical Character Recognition (OCR) technology, as evidenced in the analysis of Cherokee language preservation efforts, where the refinement of image quality yielded notable reductions in Word Error Rate (WER) and Character Error Rate (CER). Similarly, investigations into multilingual neural machine translation illuminate the superiority of Transformer models over LSTM architectures, particularly in contexts where linguistic resources are constrained. Moreover, methodological refinements, such as iterative training regimens and the integration of back-translation techniques, yield substantive improvements in translation quality, as evidenced by notable enhancements in BLEU scores across diverse language pairs. For instance, in the examination of Dravidian languages, the amalgamation of multilingual modeling and back-translation affords unprecedented gains in translation efficacy, underscoring the indispensable role of contextual information and iterative refinement mechanisms in enhancing translation fidelity. In sum, the empirical findings underscore the imperative for continued scholarly endeavor and methodological innovation in the realm of NLP. By elucidating the efficacies of tailored methodologies and empirically substantiating their efficacy through rigorous experimentation, this research paves the way for future advancements in machine translation and language processing technologies, thereby fostering cross-cultural discourse and the preservation of linguistic diversity on a global scale.

REFERENCES

[1] Khan, N.S., Abid, A. & Abid, K. A Novel Natural Language Processing (NLP)–Based Machine Translation Model for English to Pakistan Sign Language Translation. Cogn Comput 12, 748–765 (2020).
[2] Lakew, S. M., Federico, M., Negri, M., & Turchi, M. (2018). Emerging Topics at the Fourth Italian Conference on Computational Linguistics (Part 1): Multilingual Neural Machine Translation for Low-Resource Languages. Italian Journal of Computational Linguistics, 4-1, 11-25.
[3] McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in Translation: Contextualized Word Vectors. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 30). Curran Associates, Inc.
[4] Reddy, M. V., & Hanumanthappa, M. (2012, December 31). NLP Challenges for Machine Translation from English to Indian Languages. ePrints@Bangalore University. Department of Computer Science and Applications, Jnanabharathi Campus, Bangalore University, Bangalore, India.
[5] Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.
[6] Vyawahare, Aditya & Tangsali, Rahul & Mandke, Aditya & Litake, Onkar & Kadam, Dipali. (2022). PICT@DravidianLangTech-ACL2022: Neural Machine Translation On Dravidian Languages.
[7] Xie, W. (2021). Multilingual Neural Machine Translation and Back-Translation. In Proceedings of the Workshop on Dravidian Language Technology (EACL 2021). Beijing Language and Culture University, China



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY