# **Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages**

In 2022, A. Mhaske et al. developed Naamapadam which is the largest dataset for NER i.e. Named Entity Recognition. Prior to these most Indic languages don not have sufficient labeled data for NER data to build an enhanced model. The dataset resolved this problem providing us with enrich training and testing datasets. The dataset is currently available in 11 of the most prominent Indian Languages belonging from two different families. Comprising of more than 5.7M sentences the dataset is developed with 9.4M entities for basically three categories. Furthermore, a specific test dataset was created for 9 different Indic languages. The training dataset was prepared using the help of the projection approach resulting in inexpensive dataset generation. Finally, IndicBERT was developed accomplishing F1 score of more than 80 for 7 out of 9 languages. This research addresses NER training challenges in moderately-resourced languages with limited parallel data by employing span-based annotation projection to alleviate word alignment errors, particularly when dealing with low-resource languages.

Mhaske, A., Kedia, H., Doddapaneni, S., Khapra, M. M., Kumar, P., Murthy V, R., & Kunchukuttan, A. (2022). Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages. *arXiv preprint arXiv:2212.10168*.

# **Machine Translation Approaches and Survey for Indian Languages**

In 2017 N.J.Khan et al. , publish a study on Machine Translation for the Indian languages. The paper suggest use of Statistical Machine Translation(SMT) with Phrase based approach. The model focused on translating the Indic languages into English working on EMILLE (Enabling Minority Language Engineering) corpus which consist of 63 million words corpus from various Indic Languages. The model used for translation accepts a complete sentence in Indic language as an input and brakes down the sentence into different phrases. Eight commonly spoken Indian languages were selected for the project, test corpus for which were available in the dataset. BELU score were used as the parameter of evaluation. The model achieved a mean score of 0.12 and a Standard Deviation of 0.06 on the test cases provided while using 5-fold cross validation method. Despite the efforts to address translation challenges in less-explored Indian languages, the inherent divergence among source languages and the limited availability of parallel data have resulted in significant disparities in machine translation evaluation scores between seen and unseen datasets, highlighting a notable disadvantage in the current approach.

Khan, N. J., Anwar, W., & Durrani, N. (2017). Machine translation approaches and survey for Indian languages. *arXiv preprint arXiv:1701.04290*.

# **A Reference Based Machine Translation Evaluation Metric Using Linguistic Knowledge and Contextual Embeddings**

In 2023 N Joshi, and P. Katyayan developed an evaluation system for Machine Translation which could be used for linguistic as well as word embedding outputs. Performance of the same was evaluated on various English to Indic model using an adequate environmental setup. Prebuild stemmer and word embedder was used throughout the process available on open sources. REKHA(Robust Evaluation through Knowledge Harnessing Approach) was developed as a result of the comprehensive study which evaluates the performance based on different levels like lexical level, syntactical level and semantic level. The result were then evaluated against human annotator and a reasonable resembling output were yielding an faster evaluation metric approach. However the metric was lacking in the area of analyzing paraphrase and is limited to certain number of languages.

Joshi, N., & Katyayan, P. (2023, May). Rekha: A Reference Based Machine Translation Evaluation Metric Using Linguistic Knowledge and Contextual Embeddings. In *ICIDSSD 2022: Proceedings of the 3rd International Conference on ICT for Digital, Smart, and Sustainable Development, ICIDSSD 2022, 24-25 March 2022, New Delhi, India* (p. 158). European Alliance for Innovation.

# **Human Versus Automatic Evaluation of NMT for Low-Resource Indian Language**

In 2022 G. Datta et al. compared the human evaluation against parametric evaluation on Neural Machine Translation and transcribed the results. Firstly, Bengali to English corpus was selected for the project and NMT was trained over the dataset. Parallel to it the text corpus was also translated using the Google and Bing translator. Finally, the corpus was converted with the help of human annotator which is the best translation method. BLEU and WER was used as the evaluation metrics and average of 90% , 0.12 and 66% scores were obtained in the field of Human evaluation, BLEU and WER. It was concluded that fine tuning of the model enhances its performance. Hower the are of semantic evaluation still remained unexploited encouraging further research and analysis in the area of MT

Datta, G., Joshi, N., & Gupta, K. (2022, May). Human Versus Automatic Evaluation of NMT for Low-Resource Indian Language. In *The International Conference on Recent Innovations in Computing* (pp. 715-725). Singapore: Springer Nature Singapore.