**NLP PROJECT REPORT**

**MEMBERS:**

HARSHITH G

KESHAV M

VARSHITH

PRANEETH KUMAR

**ABSTRACT:**

This particular project attempts to investigate the natural language processing (NLP) techniques to be used for improving the quality of Hindi to English machine translation in terms of fluency and accuracy. Owing to the fact that Hindi and English bear a great difference in both structure and syntax, it becomes apparent those high-quality translations cannot be for free from factors such as word-order imbalance, grammar misalignment, background culture, among other possible suppressing elements. The current study focuses on the application of different techniques including neural machine translation (NMT), attention to translation, and transformer architectures, such as BERT and GPT, in translation to enhance context and idioms in translated text. In addition, the project uses pre-collected bilingual corpora and subject-related dictionaries to solve the problems of the lack of appropriate vocabulary and to enhance translation quality within various fields, including news, literature, spoken language, etc. This research shows the advantages and drawbacks in Hind-to-English translation under the deep learning architectures applied to existing NLP translations. Therefore, the developments are recommended so as to improve the flexibly and precision of these systems. This is anticipated to assist in creating tools that will be able to provide translation services to Hindi-speaking users in conversations that are more humane and contextually appropriate.

**INTRODUCTION:**

As the world has become a global village, situations that may arise in different regions are reported in a language different from one’s mother tongue and in more than one foreign language. However, the very first step of news comprehension goes further than only understanding the import, but also poses a problem since one may be overwhelmed by the vast sea of news owing to the many available genres and the various languages involved, consuming news calls for a perfect translation of the news into the audience’s language and further a summary of the most appropriate distortion of the relevant information. This project helps to meet these requirements through the use of advanced AI by translating and summarizing Hindi news articles into English with the help of intelligent systems and making information more widespread.

In order to achieve the desired result in translation of text from Hindi to English in the context of language of news articles involving low resource languages, mBART has been used which is a very powerful multilingual model designed particularly for translation. To help improve translation quality, mBART exploits the structural and functional levels of Hindi, intricacies that most languages do without. In Summarization, we employ PEGASUS, one of the state-of-the-art text summarization models as it excels in compressing large volumes of text by identifying and presenting the core information in a structured way. PEGASUS has a slightly different training methodology that involves mask-sentence training to produce a summary of quite a high quality and in a logical order as the training synthesis of the original work is preserved.

**LITERATURE REVIEW:**

**1. Multilingual Translation with Extensible Multilingual Pretraining and Fine-tuning**

In this research work, we present a model mBART, which aims to extend the scope of BART (Back and Rethink Translations) for better performance in machine translation than other models, mBART is a multilanguage BART model. mBART is introduced to support translation as well as general text generation. It does so by extending the denoising autoencoder training of BART to any multilingual data, which also works well even for those language pairs like Hindi-English that are not well resourced.

Key Components:

- Pretraining in multiple languages: mBART is pretrained on a largescale multilingual corpus with a specific denoising task, in which the model is required to recover “noisy” input text. The pretraining in many languages helps with this as the model learns to appreciate the similarities in language patterns helping it to be useful for translation activities in different language pairs.

- Fine-tuning on Specific Language Pairs: Following the pretraining phase, mBART typically goes through fine-tuning for particular language pairs, which allows it to adjust to the specifics of each language pair's style and lexicon. Thus, this enhances the quality of translation not only quantitatively, but also qualitatively in terms of the overall meaning conveyed.

- Performance: mBART demonstrated higher performance than traditional architectures such as Marian MT and the normal transformer with BLEU scores of up to 35.8 recorded in Hindi to English language translation. These figures are indicative of a higher level of correctness, indicating that mBART performs well in terms of context retention and coherence as well.

- Challenges: The major drawback of mBART is due to its architecture which is that of a transformer, which has a high computational cost requiring more resources and heavier datasets which renders it expensive to train and use especially for large multilingual tasks.

**2. PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization)**

This research presents a new training framework called PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization), a model for abstractive text summarization that aims to be the best among existing ones using an original pre-training technique called Gap-Sentence Generation. PEGASUS was created in Google Research and its main aim is to produce good summaries automatically from the available amount of data not necessarily including labelled summaries, which provides insensitivity to a concrete field.

Main Parts:

Gap-Sentence Generation (GSG) Pre-Training: In pre-training phase, PEGASUS masks out certain ‘gap sentences’ (those sentences which bear the most significance in re-constructing the original input text) making the model distort the input document for re-submission. The model is then required to reconstruct the given content by filling in the masked sentences practised in the system enabling the model acquire skills of writing short and meaningful dictations. This technique is rephrased as the summarization paradigm and is less distant in its essence to how a human constructs a summary than simple token level masking of the input.

Domain Flexibility: Since it learns on gap sentences instead of real human abstracts. PEGASUS works well in every domain such as news, scientific articles, and even medical texts. Thus, it is capable of producing summaries without extensive overfitting on the data from that particular domain.

Performance: PEGASUS demonstrates excellent performance for many abstractive summarization benchmark datasets, such as CNN/Daily Mail and XSum, and WikiHow. Its derived summaries are also consistent with each of the model, however, in terms of covering the content most salient points, this one received the highest rank in the evaluated ROUGE scores.

Challenges: One major drawback which may be popularly identified with PEGASUS is the need to have several computational resources before any form of pre-training can be done.

**3. Implementing deep learning-based approaches for article summarization in Indian languages**

This work investigates automatic summarization in the context of diverse Indian languages using various deep learning approaches owing to the factors of linguistic, resource, and syntactic challenges as in Indian languages such as Hindi, Tamil and Bengali. Many Indian Languages do not have rich infrastructures such as large corpora of tagged text which quite limits the scope of the study hence focus on both integration of existing models and onset of techniques appropriate to the peculiar language structure wisely.

Key Components:

Model Selection and Adaptation: The paper analyses some advanced models for summarization methods; in particular, sequence-to-sequence (Seq2Seq) models, transformers, and attention-based models. This work has been elaborated and extended in the case of Indian languages, in many cases it is also necessary to modify the models in order to manage the agglutination, morphology and syntactic features present in those languages.

Transfer Learning and Multilingual Modelling: Since there is a limited amount of labelled data, the interdisciplinary approach of transfer learning and the use of language agnostic models, for instance, mBART and mT5 which allow for language transfer from especially English which is a high resource language provides solutions to these problems as discussed in the paper. These models have been pretrained on multilingual corpora and adapted on very few Indian languages datasets improving summarization even with scarce data.

Evaluation and Benchmarking: ROUGE measures for assessing the quality of the summaries produced in the research are used in the study and it is evident that modified models perform on par with other approaches. Though, it should be noted that evaluation metrics must be qualified since in most cases the geometry, as exhibited, is very specific to the Indian ethnic languages, and thus existing metrics of evaluations might not be conclusive in measuring the quality of the summaries produced.

Challenges: This research also emphasizes such constraints as limited availability of annotated datasets, high processing power needed as well as language specific traits.

**PROPOSED METHODOLOGY:**

The approach presented integrates an mBART based multilingual translation engine and PEGASUS based abstractive text summarizer to build a complete system for translating and summarizing English articles. This approach combines, the multilingual translation capabilities of the mBART aimed at article translation and the summarization ability of PEGASUS, to promote effective and contextual cross lingual article summarization. The inclusions below provide the discussions on the data preps, model adaptation, pipeline integration, and evaluation procedures

1. Data Collection and Preprocessing

•Utilized Data Sources: Obtain a well-sized and a bilingual dataset from Indian language news portals, available parallel corpora, and social media. While optimizing mBART fine-tuning modes, the dataset should also have to some degree parallel related texts in any of the targeted Indian languages like Hindi, Tamil, etc., and English so as to develop translation competence of the model. Additional immersion contents (as in news or literature) will be gathered to ensure that the model is trained for more everyday applications.

•Data Usability: Setting aside certain boundaries, Indian languages pose a unique problem of diacritics, convolutional grammar, and dependency-driven morphology. Preprocessing activities include AdTech risk mitigation measures like tokenization, normalizing diacritical marks, as well as addressing localization issues such as inverted commas or cursive writing (Hindi written in English alphabets). Preprocessing also comprises dividing texts into meaningful units known as sentences, erasing unwanted symbols and special characters, formatted differently from the target language and aiming to adjust them in order to advance the training of the model decreasing distortion.

•Addressing Patterns of Use in Particular Fields: Establish and eliminate motivational, region-based phrases, idioms, or representatives of actual speech distinct to news and social media, fiction, etc. in order to increase the relevance and the ability of the text summaries to convey the required information. As political terminology and phraseology often do, these may be extremely difficult to translate or summarize without prior from normalization.

2. Translation with mBART

•Model Selection: mBART (Multilingual Bidirectional and Auto-Regressive Transformers) is Facebook AI’s in-house transformation-based multilingual model trained on tremendous amounts of data for low-resource geology. It is particularly well suited for Indian languages due to its ability to understand the similarities and variations between two languages.

•Fine-tuning on Indian Languages: Fine-tuning mBART with parallel Hindi-English or other Indian language-English datasets allows the model to specialize in certain language pairs and domains, acclimatizing to the structure and morphology of Indian languages. This step is significant since it ensures that the model better translates the two languages without losing the context and meaning of the original article.

•Handling Low-Resource Language Challenges: mBART's multilingual pretraining is one of the most important features that allow the model to perform adequately even in cases where labelled data is sparse for the target data. The model has been trained for instance on Indian languages such as Kannada and Bengali effortlessly even without extensive targeted resources for those languages.

•Evaluation of Translation Quality: The evaluation of translation performance is based on so-called BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores, with higher values being indicative of better retention of the original article’s meaning and context. Readability, coherence, and cultural appropriateness can also be assessed through human evaluation.

3. Summarization with PEGASUS

• Gap-Sentence Generation (GSG) Pre-training: GSG in terms of PEGASUS (which stands for Pre-training with Extracted Gap-Sentences for Abstractive Summarization) introduces an implementation in which quite a number of full sentences strategically selected for their ability to summarize, are cut out of the original text. This method trains the model to fill in gaps in sentence construction aided by other texts provided, such as in a form of plain answering of questions without referencing external materials allowing creation of full and meaningful summaries within short turn-around time periods as is the case of PEGASUS that does framing of its summaries within critical issues.

• Fine-tuning on Summarization Data: The last stage consists of a refinement of compatible translating techniques using a bilingual corpus. During this stage, every effort is made to modify the corpus to transcend the limits of the source-target language pair and article genres. For instance, adjusting the politics or economics news license allows importance of the jargons and formations to be captured, which enhances the relevance of these summaries in relation to the subject matter.

• Abstractive Summarization: In this instance, PEGASUS is able to create an execute an abstract by way of synthesis whereby all the key information with regard to the insights the prepare exists in a informative and precise write up. Such an efficient structure ensures that all the important facts are conveyed without excessive information, which is essential for the consumption of international news.

**EXPERIMENTAL SETUP:**

This section presents the information regarding the apparatus of the constructs, the settings, and the assessment procedures for building and testing the translating and summarizing tool based on mBART and PEGASUS technologies.

1. Data Preparation

Dataset Selection: The mBART model was well-tuned to translate Indian languages into English through this use of bilingual corpuses. However, for the case of summarization, English articles accompanied by human pattern summaries were used for the training of the PEGASUS.

Preprocessing: Text was divided into squares without dirt and regularized infusion of dirty parts, diacritics and transliteration. In addition to that, sentences were segmented in order to attain desirable form of the text based on purpose, either for translation or for summarization.

Data Splits: The data was divided into training (80%), validation (10%) and testing subsets (10%).

2. Model Configuration

mBART Translation: In order to achieve the targeted translations, the mBART model has been fine tuned on the orientation with specific learning pair of Indian languages and English.

PEGASUS Summarisation: After being fine-tuned on English translated texts, PEGASUS was tasked with generating non-specific self-summaries of the texts of the translations.

3. Computational Environment

Experiments were carried out on GPUs because they are a requirement when it comes to mBART and PEGASUS model due to the large-scale computation. The Batch size and the learning rate were set differently for the two models.

4. Evaluation Metrics

Translation Quality: BLEU scores were used to assess the accuracy of the translation.

Summarization Quality: Posterior metrics evaluated ROUGE scores and its precision, recall, and F1 of final summaries with additional information from native judges on those summaries clarity and coherence.

This arrangement made it possible to carry out all the necessary operations within a short time frame thus enabling us to perform numerous iterations of refining and testing Indian language translation and summarization models.

**RESULTS & ANALYSIS:**

BLEU Scores for:

Translation =

Summarization = 30.12

The above BLEU Scores are very good enough for the working model of this project (Translation and Summarization). If we can change any model and try some other methods, we may improve our model and make it future proof.

**CONCLUSION & FUTURE WORK:**

The current study outlines a thorough approach for translating and recapping articles from Indian languages using mBART and PEGASUS. The two techniques are standardized for the particular language models. Thus, the system can translate and summarize Hindi, Tamil, or Bengali content into English, thereby enhancing the accessibility of information across different languages. This approach tackles the problems associated with different languages, the insufficient amount of data, and the structurally and semantically complex Indian languages.

The functions of mBART in multilingual translation and of PEGASUS in summarisation show great potential in helping users comprehend and navigate content across languages. Given specific in-domain dataset for fine-tuning and resource management techniques to work with limitations of an optimal model, the proposed pipeline is distinguished by the capability to produce an appropriate translation and summarization. The evaluation results which yielded values in the BLEU and ROUGE S method assessments indicate the system’s capacity to have a positive impact on the meaning, fluency, and relevance of the finalized summaries with a view of retaining the more important aspects of the original articles.

In this regard, this is a practical approach that is flexible enough to allow use in many areas that contain multi- language content that needs translation and summarization in real times, for example, health, news and politics among others. These cutting-edge deep learning models facilitate the removal of language barriers in access to information and information sharing across different languages.

**Future Work**

The proposed system shows a lot of promise, but there are several aspects of it that can be improved further:

1. Improving Performance of Specific Domains: Even without domain-specific adaptations, the system performs reasonably well on general datasets. However, it would be interesting to adapt models to fit into the health care, legal, or scientific articles. Fine tuning for the specific domain helps in enhancing the quality of translation and an even larger extent the quality of summarisation due to the fact that the model has been trained on specific jargon and those contexts.
2. Addressing the Problem of Multilingual Summarization: In the current development, the research work might augment the system’s ability to four or five input languages at the same time, where an article, say, for instance, a Hindi-English article is written in a mixture of many languages. In that instance, the models which would be able to translate or summarise that would be very integrated, and would improve the system dramatically.
3. Adaptation within Cultural Paradigms: Languages especially the Indian ones carry a lot of cultural aspects that contributes to the general meaning of phrases. Therefore, promising areas of research include developing the models in future to consider more of culture while translating or summarizing text to not make the result linguistically accurate only but also culturally meaningful.
4. Modified Assessment Techniques: Assessment approaches in the likes of BLEU or ROUGE that are rather conventional in use for evaluation of translating and summarizing the output are in effect, cannot be found useful to convey research on the language-context and the context of most of the Indian languages. Follow up studies might focus on developing the metrics of assessment in this respect or in more practical determination of quality appraisal by inclusion of evaluators with people assessing the quality producing the texts.
5. Scaling Up and Processing Time: In order to operate the system in question efficiently and at scale optimization steps of some sort should be taken, whether it bulb distillation, stuffing, or employing lighter designs of models, that is lighter transformer like models. This will enable the system to be able to handle wider range of texts within a shorter time period which will enabled it to be used in real time purposes such as news agencies, social media assessment, and automatic content generation.
6. Enhanced Language Coverage: The methods described above in the research deal with Hindi and other major Indian languages; however, regional languages and less popular dialects such as Assamese, Punjabi, and even Kannada, which do not have much scope in NLP, should also be introduced in the future work. Regionalizing the language would make the system even more usable and beneficial given the multilingual nature of the country.
7. Customization by End Users: Allowing users to personalize the output of summarization (for instance summary length, emphasis on specific details) might enhance the system potential for different tasks. This feature would improve user experience, as it makes it possible to provide summaries that are suited to different users.

To sum it up, this paper provides a solid basis for implementing a translation and summarisation system for Indian languages in a multimodal environment. The process proposed is quite effective in facilitating the much needed cross-linguistic communication and content sharing. As it continues to be improved and widened, the system has the capacity to be of great benefit to the users in different areas and for various purposes.