Advancing Dravidian Language Processing Beyond the Sentence Level

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computational Linguistics by Research

by

Nikhil E
2018114019
nikhil.e@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Advancing Dravidian Language Processing Beyond the Sentence Level" by **Nikhil E**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Dr. Radhika Mamidi

To my parents and friends for their constant support.

Acknowledgments

First and foremost, I would like to express my heartfelt gratitude to my advisor, Dr. Radhika Mamidi, for her indispensable mentorship, expertise and support throughout my research journey. She was immensely patient with me and lent me a lot of freedom to explore my research interests. Discussions with her have always motivated me to give my best, and I feel privileged to have had the opportunity to work under her guidance. I am also extremely grateful to Prof. Kamal Karlapalem for his helpful feedback and brilliant ideas, which expanded my horizons as a researcher and challenged me to do better.

I would like to express my sincerest gratitude to my co-authors and co-researchers, Mukund, Anshul, Aditya, and Pulkit. Collaborating and brainstorming with them has helped shape my research at every step of the way. Their unique perspectives, hard work, and dedication have been instrumental in the success of our projects. I would also like to thank the annotators, Anubhav, Mudit, Ishan, Siddhant, Kunal, Anagha, Nukit, Adith, Naren, Abhinav, and Yashaswi for their tireless efforts in the creation of our data and manual verification.

I am tremendously grateful for my close friend group, "Daddycated", without whom getting through college would have been impossible. Being surrounded by such brilliant individuals has been an invaluable gift, and all the terrific memories I have made with them will stay with me forever. I am very thankful to Anjali for her unwavering support and faith in me. I would also like to thank Aryan, Pinna, Nyquil and Hemanth for all the wonderful conversations that have enriched my life.

Lastly, but most importantly, I am eternally grateful to my parents for their unrelenting support throughout my life. They have been my pillars of strength, and they have made countless sacrifices to ensure that I have every opportunity to succeed and thrive. They have also given me the freedom to explore, to make mistakes, and to learn from them, helping me develop into the person I am today. I attribute the person I am today to their unconditional love.

As I look back on my college years, I realize that the ups and downs, the successes and failures, have all been essential for my personal growth. I look forward to carrying the lessons learned and the relationships formed into the next chapter of my life.

Abstract

The rapid advancements in natural language processing have primarily focused on high-resource languages, leaving low-resource languages, such as those in the Dravidian family, underrepresented in terms of research and resources. To bridge this gap and enable more effective processing of Dravidian languages, this thesis explores the utilization of context beyond the sentence level, delving into linguistic structures at the paragraph and document level, and leverages multilingual training. This is done particularly for the tasks of neural machine translation and multi-class sentence classification.

CoPara, the first publicly available paragraph-level n-way aligned corpus for Dravidian languages (Kannada, Malayalam, Tamil, Telugu) and English, was created. This dataset fills a critical gap in multilingual resources for low-resource Dravidian languages and enables research on the impact of paragraph-level information on NMT tasks. The corpus contains aligned paragraphs across English and four Dravidian languages, providing a rich resource for studying cross-lingual phenomena and improving machine translation quality.

To demonstrate the utility of CoPara, neural machine translation experiments were conducted by fine-tuning a pre-trained multilingual sequence-to-sequence model on the dataset. The results show significant improvements in translation quality for paragraphs across all language pairs over models trained using sentence-level data, highlighting the potential of leveraging paragraph-level information and multilingual training for enhancing machine translation systems in low-resource settings.

An annotated dataset of SEBI legal case files, along with its Indic adaptation, was created. The dataset consists of sentences classified into legally applicable categories and is the first of its kind for Indian legal adjudication orders concerning insider trading. This dataset facilitated the development of a novel system for aspect-based semantic segmentation of legal case files, tailored to different stakeholders like investors, defense lawyers, and adjudicating officers. The system leverages document-level context to improve sentence classification performance, and multilingual training using the Indic adaptation further enhances the results. The results highlight the importance of incorporating con-

text beyond the sentence level and leveraging multilingual training to advance Dravidian language processing.

In summary, the thesis provides a comprehensive exploration of utilizing paragraph and document-level context for NMT and sentence classification tasks. The datasets and methodologies introduced pave the way for future research on long-text processing and context-aware NLP tasks in low-resource languages, with potential implications for other language families and domains.

Contents

List of Figures

List of Tables

Chapter *1*

Introduction

The objective of this thesis is to advance Indic language processing beyond the sentence level by exploring linguistic structures at the paragraph and document level, addressing the challenges and limitations of existing resources and techniques in processing longer textual forms for low-resource Dravidian languages.

1.1 Motivation

Incorporating context beyond the sentence level and leveraging multilingual training are two crucial approaches that have gained significant attention in the field of natural language processing (NLP). These techniques have been shown to improve the performance of various NLP tasks, such as **neural machine translation** (NMT) and **multi-class classification**. The motivation behind these approaches lies in their ability to capture valuable information that is often lost when processing text at the sentence level alone.

By considering larger units of text, such as paragraphs or entire documents, NLP models can capture important discourse-level features. Exploring such structures could significantly increase the capabilities of evolving models, allowing them to efficiently process longer textual forms with greater accuracy and comprehension[94]. These features include coherence, which refers to the logical flow and consistency of ideas within a text, and cohesion, which relates to the linguistic devices used to connect sentences and create a unified whole. Additionally, incorporating broader context allows models to understand better the rhetorical structure of a document. This exploration holds the potential to address various linguistic challenges, ranging from basic tasks such as paragraph translation to more complex tasks such as coreference resolution and author style differentiation[26]. Contextual information also helps in resolving ambiguities that arise

due to the lack of surrounding context at the sentence level. By leveraging the additional information provided by the larger context, models can disambiguate word senses and make more accurate predictions overall. Furthermore, certain linguistic phenomena, such as anaphora resolution (identifying the antecedent of a pronoun) and the interpretation of discourse markers, rely on information that spans multiple sentences. **Incorporating context beyond the sentence level** enables NLP models to handle these long-range dependencies better and improve the overall quality of the output.

Multilingual training, on the other hand, offers several benefits that can significantly enhance the performance of NLP models. By training on multiple languages simultaneously, models can learn shared representations that capture language-agnostic features. This enables transfer learning from high-resource languages to low-resource languages, which is particularly beneficial for tasks like NMT where parallel data is scarce for many language pairs. Multilingual training also promotes improved generalization, as exposure to diverse languages helps models learn more robust and generalized representations. This leads to better performance on unseen data and improved handling of linguistic variations within and across languages. Moreover, multilingual models are less prone to overfitting and can better capture the underlying linguistic structures shared across languages. From a practical standpoint, multilingual training allows for more efficient utilization of computational resources and data. Instead of training separate models for each language pair or task, a single multilingual model can handle multiple languages and tasks simultaneously, reducing overall training time, storage requirements, and maintenance efforts. Additionally, multilingual models enable zero-shot or few-shot learning, where the model can perform tasks in languages it has not explicitly seen during training by leveraging the shared representations learned from other languages.

Relevant datasets are the backbone of exploring these approaches. To effectively incorporate context, datasets should provide longer units of text, such as paragraphs or entire documents, along with annotations for discourse-level features, rhetorical structure, and instances of ambiguity resolution, coreference resolution, and anaphora resolution. Similarly, for multilingual training, diverse datasets covering a wide range of languages are essential. Parallel corpora, which align texts across multiple languages, are particularly valuable for tasks like neural machine translation, enabling the learning of shared representations and facilitating transfer learning between languages. Moreover, datasets that include linguistic variations within and across languages are crucial for promoting generalization and robustness in multilingual models, ultimately enhancing the performance and efficiency of NLP systems. The combination of longer text units and multilingual parallel data has the potential to significantly improve NLP performance across various

tasks and languages. Models trained on such datasets can better capture complex linguistic phenomena, resolve ambiguities, and generate more coherent and contextually appropriate output. The increased diversity and richness of these datasets can also enhance the efficiency of NLP systems by reducing the need for language-specific training data and enabling more effective transfer learning.

1.2 Contribution

This thesis explores the use of context beyond the sentence level and multilingual training to advance low-resource Dravidian language processing, particularly for the tasks of NMT and multi-class Sentence Classification. The major contributions of the thesis are as follows:

1. CoPara, the first publicly available paragraph-level n-way aligned corpus for Dravidian languages (Kannada, Malayalam, Tamil, Telugu), and English was created. By focusing on Dravidian languages, which are relatively low-resource compared to many other language families, CoPara aims to bridge the gap in multilingual resources and facilitate research in this underrepresented linguistic area. The corpus contains aligned paragraphs across English and four Dravidian languages, providing a rich resource for studying cross-lingual phenomena and improving machine translation quality. This new dataset fills a gap in multilingual resources for low-resource Dravidian languages and enables research on the impact of paragraph-level information on NLP tasks like machine translation. Additionally, article-level and sentence-level extensions were made using the same, which can fuel further experimentation.

2. To demonstrate the utility of CoPara, neural machine translation experiments were conducted by fine-tuning a pre-trained multilingual sequence-to-sequence model on the dataset. The results show significant improvements in translation quality for paragraphs across all language pairs over models trained using sentence-level data, highlighting the potential of leveraging paragraph-level information, as well as multilingual training for enhancing machine translation systems in low-resource settings. Consequently, the importance of long-text and multilingual corpora is emphasized.

3. Creation of an annotated dataset of SEBI legal case files, with sentences classified into 10 legally applicable categories. This is the first such dataset for Indian legal

adjudication orders concerning insider trading, and was done with the help of legal experts. An Indic extension was also created by translating the dataset into four Dravidian languages, and the quality was validated both artificially and by humans. Because all sentences from the case files are used, leveraging document-level context becomes possible.

4. A novel system for aspect-based semantic segmentation of legal case files is developed, tailored to different stakeholders like investors, defense lawyers, and adjudicating officers. The system uses the SEBI legal dataset and leverages document-level context to classify sentences according to the pre-defined labels. The same process was repeated for the Dravidian extension created for the dataset, which demonstrates the performance improvements of using multilingual training in addition to document-level context. Thereafter, relevant summaries are generated by paraphrasing the persona-specific information. Thorough experimentation comparing various methods for each part of the pipeline was undertaken.

1.3 Thesis Overview

Chapter 1 provides the motivation, key contributions, and overview for the thesis on advancing Dravidian language processing beyond the sentence level with multilingual training.

Chapter 2 discusses the relevant background and prior research on paragraph-level aligned corpora, neural machine translation for Indian languages, corpora for the legal domain, using document-level context for sentence classification and summarization methods.

Chapter 3 introduces CoPara, the first publicly available paragraph-level n-way aligned corpus for Dravidian languages and English, describing its data source, creation process, statistics, quality evaluation, and a sentence-level extension.

Chapter 4 shows experiments conducted to finetune pretrained IndicBART models on CoPara for paragraph-level neural machine translation between Dravidian languages and English, demonstrating improved performance over baseline models.

Chapter 5 introduces an annotated dataset of SEBI legal case files, along with its Indic adaptation, to develop a system for aspect-based semantic segmentation of case files that leverages document-level context and multilingual training for improved sentence classification and finally personalized summarization.

Chapter 6 summarizes the main findings of the thesis and discusses potential directions for future research.

Chapter 2

Background and Related Work

2.1 Background of Dravidian Languages

A concise overview of the four Dravidian languages used for the work done within this thesis would help set the context for the subsequent chapters. The following is a summary of the morpho-syntactic characteristics and history of the four languages drawn upon the comprehensive compilation created by Gutman and Avanzati (2013) [32]. A more thorough socio-linguistic background for the relevant languages can be found in Madasamy et al. (2022) [56].

The Dravidian language family, primarily spoken in South Asia, particularly in South India, comprises over twenty languages with diverse characteristics. Four major languages - Tamil, Telugu, Kannada, and Malayalam - have developed their own scripts and rich literary traditions. These languages are spoken by millions, while others have much smaller speaker populations. Some Dravidian languages have been known for more than a millennium, while others were discovered more recently in the 19th and 20th centuries[43]. Dravidian languages have coexisted and interacted with Indo-Aryan languages, resulting in mutual influence through loanwords and phonological changes, significantly contributing to the development of Indian civilization. The Dravidian languages are divided into four groups: Northern, Central, South-Central, and Southern, with the Southern group being the most prominent numerically. The geographical distribution of these languages often coincides with state boundaries in India.

These languages, primarily agglutinative in nature, generally adhere to a head-final structure with a flexible Subject-Object-Verb (SOV) word order, requiring the finite verb to be placed at the end of each [44]. A single finite verb may be accompanied by one or more non-finite verbs within a sentence, and subordinate clauses usually precede the main

clause. **Kannada**, a pro-drop language, employs case suffixes, postpositions, participles, gerunds, and infinitives for its syntactic structure, allowing for subject omission due to the verb's ability to express person and number. In **Tamil**, subject-verb agreement is present, with verbs agreeing in person, number, and gender, and syntactical relations are indicated through postpositions and case inflection. **Telugu** expresses its syntactic functions through case suffixes and postpositions that follow the oblique stem, lacks coordinating conjunctions, and displays lengthened final vowels in coordinated phrases. Relative clauses in Telugu are formed using a relative participle. **Malayalam** closely resembles Tamil but has diverged since the 8th century, losing the agreement between subject, verb, person, and number while retaining the agglutinative nature of Dravidian languages.

2.2 Paragraph-level Multilingual Corpora

Multilingual data and NLP for Indic languages have grown in recent years, providing large parallel datasets and models [75, 2]. Dravidian NLP has also seen quality research through community efforts and workshops like DravidianLangTech [56]. While improving sentence-level NLP for Dravidian languages, structures beyond sentences should be explored using publicly available data to inform models of longer form texts [94]. This could lead to solving problems like paragraph translation and coreference resolution [26].

The first paragraph/passage level multilingual n-way aligned (PLA) dataset in Dravidian languages would be a great step towards research on document level parallel corpus creation/NMT [20, 91] and PLA corpus creation NLP [85, 94, 15, 29]. The benefits of PLA corpora have been demonstrated through the works mentioned above and classic sentence-level aligned (SLA) corpora such as Europarl [42]. These studies have showcased PLA's direct impact on improving NMT, obtaining higher-quality sentence-level aligned texts, connecting entities across Wikipedia-like databases, and enhancing literary and medical text translation. Additionally, in the context of Dravidian languages, early research by [36] and more recent findings from DravidianLangTech'21 [5] have highlighted that sentence length and complexity pose significant challenges to achieving accurate translations within this language family.

Research on paragraph-level alignment (PLA) pairs within larger second language acquisition (SLA) datasets, such as Europarl, has been conducted, but there is limited work focused exclusively on creating a PLA corpus or performing natural language processing (NLP) tasks on such data. Two notable contributions are the Par3 dataset, which is

multilingual but not n-way [85], and a Chinese-English translated novels' dataset [94]. Thai et al. [85] conclude that current metrics, such as BLEU [66], are insufficient for evaluating paragraph alignment or translation quality, emphasizing the importance of human evaluation. They also observe that neural machine translation (NMT) sentence-level translations are overly literal compared to human translations, yet BLEU favors Google Translate over human translations. Furthermore, they identify paragraphs as a crucial unit for a literary paraphrase dataset and report BLEU scores of only 15-17 for English-Tamil translation quality in the literary domain among Dravidian languages. Similarly, Zhang et al. [94] proposed an innovative hierarchical model that learns both word-level and sentence-level features to model paragraph-level units in the literary domain. They emphasized that paragraph alignment is a non-trivial task that facilitates sentence alignment while providing a richer context for NMT models.

Gao et al. [26] investigated semi-automatic methods to assess a translator's style and emphasized that paragraph-level automatic textual aligners are less prone to errors compared to sentence-level aligners in Chinese-English translations. They also highlighted that sentence boundaries are not always clearly defined by punctuation, making paragraphs a more suitable unit for alignment. Similarly, Gupta et al. [31] developed a Hindi-English aligner and observe that it performs better with aligned paragraphs. Lastly, Gottschalk and Demidova [29] demonstrated that paragraph-level alignment (PLA) with overlapping information in partner Wikipedia articles facilitates the creation of a comprehensive overview of shared entity facets across multilingual editions. These studies collectively underscore the importance and effectiveness of paragraph-level alignment in various language pairs and domains.

2.3 Long Text Neural Machine Translation

Exploring document-level or paragraph-level Neural Machine Translation (NMT) involves delving into models that consider broader contextual information beyond individual sentences. Researchers have proposed various approaches to enhance NMT by incorporating document-level context. For instance, Zhang et al.[92] introduced a method for document-level NMT based on the Transformer architecture, where previous sentences are utilized as document information, enhancing the encoder and decoder interactions through attention mechanisms. This approach aims to capture richer contextual cues for more coherent translations.

Moreover, the work by [41] emphasized the utility of document-level context in NMT experiments, where providing one previous source sentence as document-level context yielded promising results. Additionally, recent studies have focused on improving context-aware NMT models by considering full document context, as highlighted by maruf-etal-2019-selective. By exploring the integration of document-level information, researchers aim to enhance translation quality and coherence by leveraging a broader context for the NMT models. The research by [80] proposed incorporating document context into NMT training objectives, suggesting a scheme that allows for document-level evaluation metrics to be integrated into the training process. This approach underscores the importance of training NMT models to effectively utilize document-level information for more accurate and contextually appropriate translations. Overall, the exploration of document-level or paragraph-level NMT signifies a shift towards more comprehensive and context-aware translation models that can capture the nuances of language at a higher level of granularity.

To further explore the realm of document-level or paragraph-level NMT, researchers have investigated various avenues to enhance the performance and capabilities of pre-trained models. For instance, Sennrich et al.(2016)[81] delved into the utilization of monolingual data to improve NMT models without altering the neural network architecture. This approach focuses on training strategies that leverage the inherent capacity of encoder-decoder NMT architectures to learn information akin to language models, showcasing the potential for enhancing NMT through innovative training methodologies. Mansimov et al.(2020)10.48550/arxiv.2003.05259 introduced a method to capture document context within sentence-level NMT models through self-training, eliminating the need for specialized models on parallel document-level corpora. By applying this approach during decoding, the contextual understanding of models can be enhanced without the requirement of additional training data.

2.4 Pretrained Models for Indic NMT

The rise of NMT and the development of extensive parallel corpora have revolutionized the domain of Indic NMT. Initially, Recurrent Neural Networks were employed, but later, transformer-based methods[86] gained prominence. The incorporation of attention mechanisms and subword-based modelling tackled the challenges of word order and data sparsity, enabling the generation of fluent and precise translations. Several notable NMT models have been developed specifically for Indian languages, showcasing

the progress in this field[70, 22]. These advancements were followed by the emergence of multilingual and pre-trained MT models[45, 51, 13], which leverage vast amounts of training data and exploit linguistic similarities across languages. This knowledge transfer from high-resource to low-resource languages has made it feasible to develop high-quality MT systems for languages with limited resources[11]. In recent years, the availability of large corpora[75, 84, 25] and the development of more expansive models have significantly enhanced translation quality. Current research has also delved into the translation of extremely low-resource languages, which have minimal parallel corpora and limited monolingual corpora, pushing the boundaries of what is possible in the realm of Indic MT[3, 58].

2.5 Corpora for Legal Text Processsing

In recent times, the legal system in many populous countries has been inundated with a large number of legal documents and pending cases. There is an imminent need for automated systems to process legal documents and help augment legal procedures. However, the processing of legal documents is challenging and is quite different from conventional text processing tasks. Legal documents are typically quite long, highly unstructured, noisy, and use domain-specific jargon. Thus, creation of legal text processing corpora for addressing the challenges associated with the legal domain becomes important.

There are a lot of datasets available in the legal domain suitable for various NLP tasks. The CaseHold dataset[96] provides valuable insights into legal case outcomes and judicial decisions, while the Open Legal Data project[65] offers a comprehensive dataset of German court decisions. The EURLEX57K dataset[7] is a significant resource for natural language processing tasks in the legal domain. Datasets from the Lynx project[18] contribute to understanding lynx habitat use patterns and population dynamics. The Contract Discovery corpus[4] is essential for training models in contract analysis and extraction tasks, and Leitner et al. (2020)[48] present a Named Entity Recognition (NER) dataset for German. The Competition on Legal Information Extraction/Entailment (COLIEE) workshop[27] provides several datasets that are invaluable for legal text processing tasks. COLIEE datasets are carefully curated and annotated by legal experts, serving as benchmarks for evaluating the performance of various legal text processing models and fostering innovation in the field. As for Indic corpora, the Hindi Legal Documents Corpus (HLDC) [39] presents more than 900,000 documents which were cleaned and structured to be utilized for downstream NLP tasks. In addition, MILPaC

(Multilingual Indian Legal Parallel Corpus)[57], the first parallel corpus of legal text in English and nine Indian languages, including several low-resource languages, was created for evaluating Machine Translation performance in the legal domain.

2.6 Pretrained Models for Sentence Classification

Pretrained models have revolutionized the field of sentence classification, proving to be more effective than classical machine learning methods. These models, such as BERT[16] and RoBERTa[55], leverage the power of large-scale pretraining on diverse text corpora to capture complex linguistic patterns and semantic relationships within text data. By learning rich representations of language during the pretraining process, pretrained models can effectively understand the context and meaning of words and sentences, surpassing the capabilities of traditional machine-learning approaches[78, 82]. The ability to fine-tune these pretrained models for specific sentence classification tasks further enhances their performance, allowing them to adapt to the nuances of the target dataset[74]. The superior performance of pretrained transformers across various classification tasks highlights their potential to revolutionize natural language processing and pave the way for more accurate and efficient classifiers in a wide range of applications.

The availability of legal-domain-specific pre-trained language models like LEGAL-BERT[8] and LexLM-RoBERTa[9] allows further advancements in legal NLP research and applications. LEGAL-BERT is a family of BERT models pre-trained on a diverse corpus of English legal text, including legislation, court cases, and contracts from multiple jurisdictions. LexLM are RoBERTa-based models pre-trained on the newly introduced multi-jurisdictional LeXFiles legal corpus. LexLM models are warm-started from generic RoBERTa weights and further pre-trained for 1M steps on legal text using techniques like increased masking rate and exponential smoothing of sampling rates across the legal sub-corpora. Already pretrained on legal data, these models can be used to improve performance significantly on downstream tasks.

2.7 Document-level Context Incorporation

Incorporating context information has been a key focus to enhance the accuracy and performance of sentence classification models. One common approach is the use of Conditional Random Fields (CRF) to incorporate context information for classification[52].

Recent advancements have explored incorporating broader context, such as sentence and document-level representations, in addition to token representations, to improve classification accuracy. Various methods have been used to incorporate document-level context, such as hierarchical structures like Hierarchical Attention Networks that utilize a combination of sentence vectors from LSTM to form a document-level vector representation for classification[90], deep neural networks that enhance document-level sentiment classification by automatically determining the importance of sentences within documents through gate mechanisms[10], and attention mechanisms like Convolution-Based Neural Attention models that integrate context information by capturing contextualized local information within sentences[19]. Recent studies have also emphasized the significance of considering document-level context for tasks like machine translation, with document-level neural machine translation methods focusing on leveraging cross-sentence context to enhance translation performance[38] and models like Hi-Transformer aiming to improve sentence context modeling for more efficient long document processing by propagating global document context to each sentence[87]. Huang et al. (2021)[35] propose a hierarchical hybrid neural network with multi-head attention (HHNN-MHA) for document classification that leverages both sentence-level and document-level context. The model uses a CNN-BiGRU architecture with multi-head attention at the word level to capture local and global semantic features within sentences, and BiGRU with attention at the sentence level to obtain the overall document representation for classification. This approach aims to effectively incorporate both sentence and document context to enhance document classification accuracy.

2.8 Aspect-based Summarization

Aspect-based summarization is a specialized form of summarization that focuses on generating concise summaries based on specific aspects or points of interest within a given text. Hayashi et al. (2021)[34] introduced the WikiAsp dataset, which serves as a valuable resource for multi-domain aspect-based summarization tasks. This dataset facilitates research in automatic summarization by providing annotated data for training and evaluation purposes. Furthermore, Frermann Klementiev (2019)[24] addressed aspect-based summarization by developing models that generate summaries centered around specific aspects within a document. Their work highlights the importance of tailoring summaries to focus on key aspects of interest.

Zhu et al. (2009)[98] proposed aspect-based sentence segmentation as a method for sentiment summarization, emphasizing the extraction of relevant sentences associated with specific aspects to form coherent summaries. This approach underscores the significance of incorporating aspect-specific information in the summarization process to enhance the overall quality of the summaries. In the realm of opinion mining and sentiment analysis, Ravi Ravi (2015)[76] conducted a comprehensive survey that outlined various tasks, approaches, and applications in this domain. While not directly focused on aspect-based summarization, this survey provides valuable insights into sentiment analysis techniques that can be leveraged in aspect-based summarization tasks.

Chapter *3*

CoPara: The First Dravidian Paragraph-level n-way Aligned Corpus

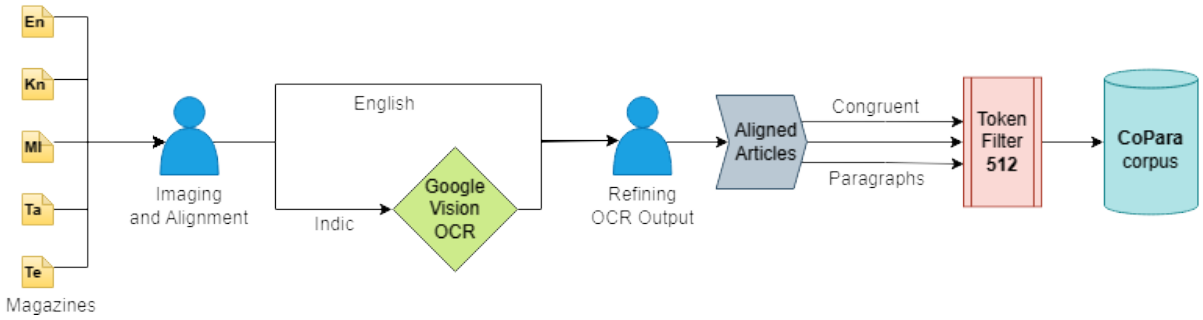


Figure 3.1: CoPara Creation Pipeline

3.1 Motivation

Parallel corpora, which consist of texts and their translations in multiple languages, have become essential resources for various natural language processing tasks, particularly in the field of machine translation. While most parallel corpora are aligned at the sentence level, there is a growing recognition of the importance of considering larger units of text, such as paragraphs, to capture discourse-level information and improve translation quality. Paragraph-level parallel corpora provide a more comprehensive context, allowing machine translation systems to better handle coherence, cohesion, and other discourse phenomena that extend beyond individual sentences. Furthermore, the development of n-way aligned paragraph-level corpora, where the same content is aligned across multiple

languages, opens up new possibilities for multilingual research and cross-lingual transfer learning.

Despite the potential benefits of paragraph-level parallel corpora, their availability remains limited, especially for low-resource languages. Many language families, such as Dravidian languages, have been underrepresented in the realm of multilingual resources, hindering the progress of research and the development of effective machine translation systems for these languages. The creation of a high-quality, n-way aligned paragraph-level corpus for Dravidian languages would not only fill this gap but also enable researchers to explore the impact of paragraph-level information on machine translation quality in low-resource settings. By leveraging such a corpus, pre-trained multilingual models can be fine-tuned and the extent of translation performance through paragraph-level context can be investigated, paving the way for more accurate and fluent translations in Dravidian languages and potentially other low-resource language families.

CoPara, the first publicly available paragraph-level n-way aligned corpus for Dravidian languages, aims to address the need for multilingual resources that go beyond the sentence level. By focusing on Dravidian languages, which have been underrepresented in the field of parallel corpora, CoPara seeks to fill the gap in resources and enable researchers to explore the impact of paragraph-level information on machine translation quality in low-resource settings.

3.2 Data Source

The dataset is derived from the "***New India Samachar***" magazine [71], a publication launched by the Information and Broadcasting (I&B) Ministry of India in 2020. The publication cycle of the magazine is fortnightly and it is available in English and twelve different Indic languages. Among these twelve languages, four are Dravidian languages, namely Kannada, Malayalam, Tamil, and Telugu. The magazine disseminates information on various cabinet decisions, features content like 'Mann ki Baat', a radio program hosted by the Prime Minister of India, and discusses prevailing issues. The parallel texts collected cover a large range of topics in the magazine news domain, such as politics, economy, culture, and social issues. Using these texts for the dataset can help improve the quality and fluency of machine translation and enable cross-lingual analysis and understanding of the magazine news domain.

The magazines contain many images wrapped in text, and the content is often arranged in multiple columns within a page. Each issue of the magazine contains multiple articles

about varying topics, and each article is structured into a heading (and sometimes even a sub-heading) and multiple paragraphs/bullet points. Each of these articles sometimes spans multiple pages and often contains multiple sidebars that detail information either complementing or supplementing the main material within the article but can also act as standalone articles on their own. The same articles span the same pages and have very similar formatting across all language versions of the same issue (As shown in Figure 3.2), but the number of paragraphs/bullets that make up these articles and how the same text is arranged within the page often differs across versions. Additionally, the text present in the magazines is either in image format or non-standard encoding, rendering automatic extraction of the required text close to impossible.

(a) English

(b) Telugu

Figure 3.2: Identical corresponding pages for the same issue in different languages

3.3 Dataset Description

15 issues published in 2022 were processed (Pipeline at Figure 3.1) to create *CoPara*, with a subset picked randomly. A total of 75 issues were targeted for alignment in an n-way manner across 5 languages: English and the four Dravidian languages (Kannada, Tamil, Telugu, and Malayalam). This process resulted in **2856** n-way aligned paragraphs and **1293** n-way aligned articles, with statistics highlighted in Tables 3.1 and 3.2. These statistics demonstrate that *CoPara* aligns with general relative linguistic features. For instance, each sentence that is part of an n-way aligned paragraph in English will, on average, be longer than a sentence in a Dravidian language when considering Word Length, as Dravidian languages are more agglutinative.

Avg. Len.	**kn**	**ml**	**ta**	**te**	***en***
Tokens	100.6	105.4	103.2	105.8	*108.6*
Words	49.7	45.8	53.0	52.1	*70.2*
Sentences	3.9	4.1	4.3	4.5	*3.8*

Table 3.1: Average Lengths of *CoPara* paragraphs on various levels of units across languages

Avg. Len.	**kn**	**ml**	**ta**	**te**	***en***
Tokens	188.6	197.2	193.1	198.6	*199.8*
Words	93.4	86.3	99.7	97.9	*132.9*
Sentences	7.2	7.7	8.1	8.4	*7.1*

Table 3.2: Average Lengths of articles on various levels of units across languages

The following sections detail the steps of the data processing and show a detailed analysis of alignment quality as well.

3.4 Data Creation

3.4.1 Utilizing OCR

Given the characteristic presence of image-based text or non-standard encoding in all the Dravidian language magazines (because they are in PDF format), direct text extraction was highly erroneous. This necessitates using standard Optical Character Recognition (OCR) software for more accurate text extraction. **Google Cloud Vision Document Text Detection**, Amazon Textract (API-based proprietary solutions provided by Google and AWS respectively), and Tesseract OCR (open source solution first developed by Hewlett-Packard in the 1980s) were experimented with, but the first was chosen because of high-precision image-to-text conversion for Dravidian languages, optimization for dense text, and support for code-mixed output which is necessitated by the presence of some named entities in the Dravidian versions of the magazines. In addition, the Cloud Vision API also leverages page, block, paragraph, word, and break information for dense documents. As the content is often present in multiple columns and the articles contain varying numbers of components across different versions, the additional information about the text positioning processed by this solution is extremely instrumental for this specific use case. The generated output contains the text itself and the *bounding box* information for the various hierarchical elements within the PDF in JSON format.

3.4.2 Organization and Alignment

Despite all the positional information that the OCR solution chosen can process and output, in many instances, a single article is segmented by images and sidebars across pages, and the columns within the page are not always in order of content. Consequently, the sequence in which the text is outputted is affected even by minor variances in relative positioning and indentation. Therefore, the OCR output is unreliable for obtaining correctly organized text because of the unpredictable order in which the solution chooses bounding boxes. In addition, the solution's ability to keep the order of text the same across outputs for the various language versions is even worse because of even more variations in the formatting and arrangement. Hence, depending on the naturally produced outputs of the OCR solution, for both the organization of data within a language and the alignment of data points across languages is not feasible. This requires manual organization and alignment of the data points obtained.

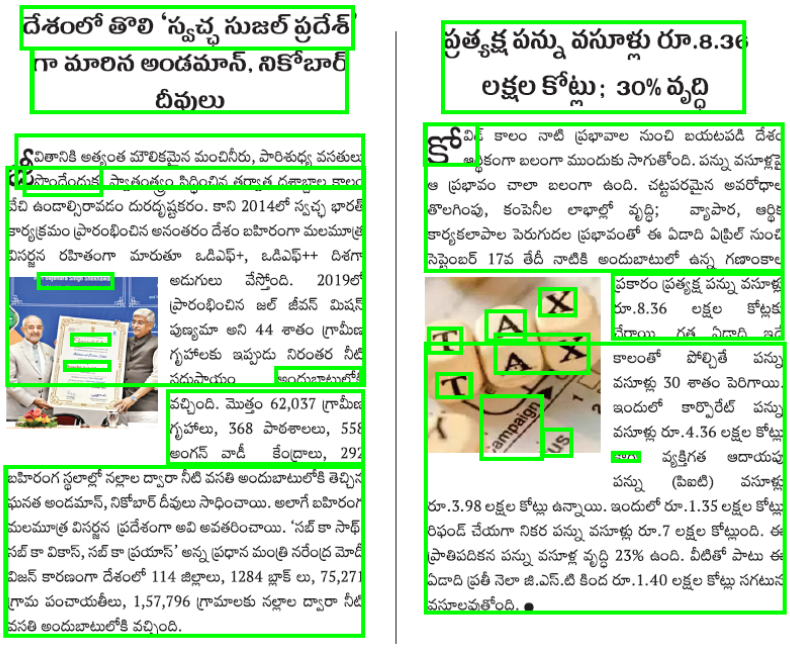
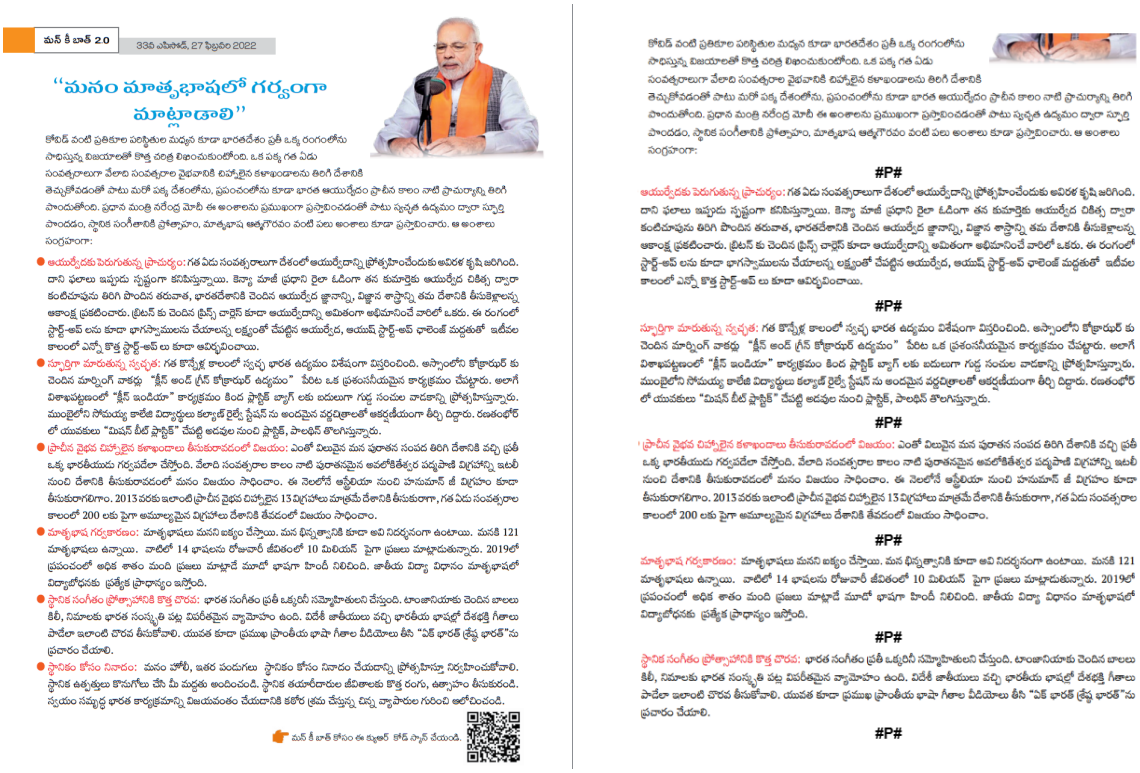


Figure 3.3: Inconsistencies within paragraph level bounding box information

One possible way to leverage the bounding box information is to automatically annotate the rectangles specific to the paragraphs using the *Fitz Python library* in the PDF files. Thereafter, screenshots can be generated using the annotations within the files. Finally, they can be organized correctly and aligned across languages manually. However, this approach fails as the bounding box information generated for paragraphs is severely inaccurate because of inconsistent indentation within/between blocks of text and varying font sizes (an example is shown in Figure 3.3). To compensate for this inaccuracy, the procedure of creating images for the various paragraphs must also be manual.

Six annotators were employed for the final methodology of organization and alignment chosen. The English magazine content is uniquely text-based with standard encoding instead of the image-based and uninterpretable nature of all other language versions. The annotators organized this content into articles with varying numbers of paragraphs after being briefed on the methodology of using visual cues like relative positioning, spatial heuristics, and matching design elements. Subsequently, the same organization is annotated within the Dravidian language versions of the corresponding magazine containing the corresponding articles, which may or may not have the same number of paragraphs. These annotations are then used to capture manual screenshots of the required blocks of text from the magazines. The screenshots are organized into text documents, which are segmented by indicating breaks throughout the entire content using a combination

of **article breaks** (*A*) and **paragraph breaks** (*#P#*), as shown in Figure 3.4. This was then checked for errors and re-annotated if required until it was satisfactorily aligned. Upon completion, there exists a set of five image documents corresponding to each of the five language versions being used for each issue of the magazine chosen. Each of the documents in a set contains the same number of articles, but differs from the others in the total number of paragraphs.



(a) Sample original page

(b) Image document for the same page

Figure 3.4: A sample page of the magazine before and after annotation

3.4.3 OCR and cleaning

The image documents generated in the previous step are used with the OCR solution chosen to generate the corresponding raw text. The same team of annotators then refined the generated textual data to maintain quality. Through comparison with the image documents, they denoised the image-text conversion by tackling issues such as misinterpreted characters, incorrect order, missing words, inappropriate formatting, and other noisy artefacts. Following this refinement, the text is subjected to further standard

text pre-processing to remove extra punctuations, redundant whitespaces, line breaks, and hyperlink fixes.

3.4.4 Splitting into paragraphs

The image documents obtained from the alignment phase are broken down so the OCR solution can reliably reproduce these breaks in the final output. These breaks serve as delimiters, simplifying the conversion of the processed text output into the dataset format. Utilizing the breaks in the articles, the text is subsequently aligned, yielding 1893 articles aligned in an n-way manner. To guarantee precise paragraph alignment for the subsequent step, each article in all languages is scrutinized, with the number of internal paragraphs being tallied. In cases where the count is the same, the corresponding paragraphs are presumed congruent, and the tuple is included in the dataset (consistent with the assumptions of prior work). This remains true even if the count amounts to a single paragraph.

In the final step, each paragraph from all tuples (which are now aligned) is tokenized using the **IndicBART-XXEN**[1] [13] tokenizer from the Hugging Face Library. If the resulting token count exceeds 512, the corresponding tuple is excluded to ensure greater compatibility with existing language models. This meticulous process results in a parallel paragraph-level corpus comprising **2856** passages aligned n-way, where n equals five (English, Kannada, Tamil, Telugu, and Malayalam).

3.5 Data Quality

The methodology from prior research on quality estimation for multilingual corpora of parallel sentences in Indic languages [75] has been modified to suit the task of estimating quality at the paragraph level. A pair of experiments were conducted to estimate the Semantic Textual Similarity (STS) of the data tuples, which indirectly measure their alignment quality, given that a correctly aligned paragraph would typically have high similarity with its counterpart in another language.

These two experiments were carried out on a subset of the data that was randomized and balanced in length (5%). The first experiment involved an Artificial estimate (Section 3.5.1), where the cosine similarity of cross-lingual embeddings was calculated. The second experiment involved a Human estimate (Section 3.5.2), where human annotators were

[1] https://huggingface.co/ai4bharat/IndicBART-XXEN

asked to rate the same sentences on a scale of 0-5, specifically designed for an STS (semantic textual similarity) task.

3.5.1 Artificial Alignment Evaluation

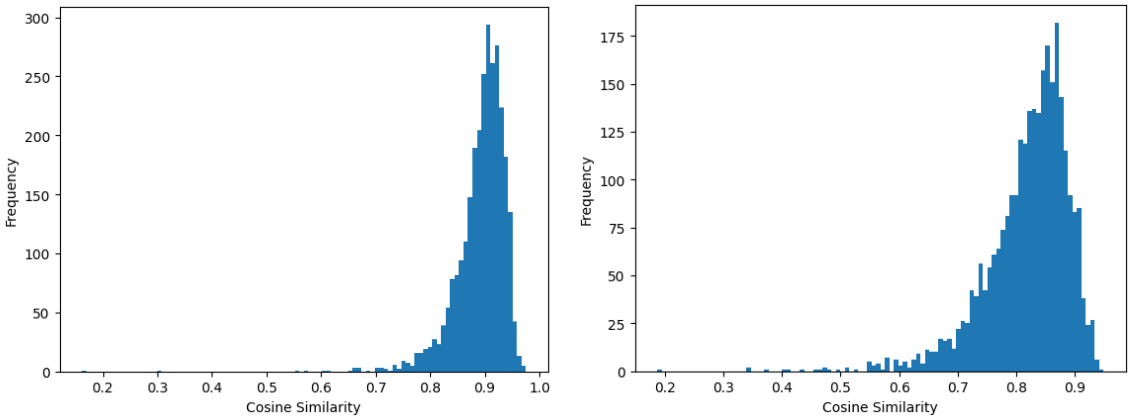
statistic	**en-kn**	**en-ml**	**en-ta**	**en-te**
Mean	0.892	0.819	0.864	0.852
SD	0.047	0.073	0.065	0.064
Min	0.162	0.187	0.137	0.218
Max	0.975	0.948	0.979	0.965

statistic	**kn-ml**	**kn-ta**	**kn-te**	**ml-ta**	**ml-te**	**ta-te**
Mean	0.842	0.876	0.872	0.827	0.821	0.850
SD	0.069	0.062	0.061	0.078	0.073	0.072
Min	0.361	0.271	0.290	0.253	0.300	0.237
Max	0.958	0.977	0.977	0.965	0.947	.976

Table 3.3: Descriptive statistics for cosine similarity scores for measuring alignment between all language pairs

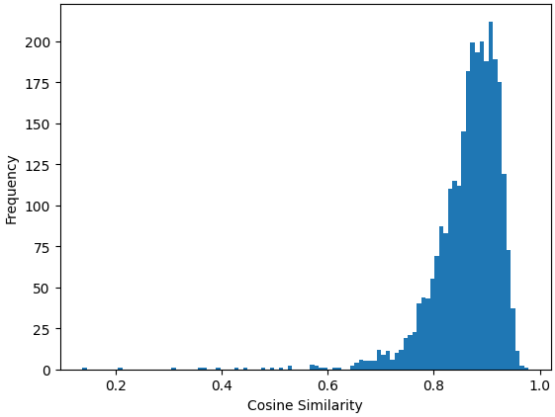
The alignment of data tuples was artificially assessed by generating cross-lingual embeddings with the **indic-sentence-similarity-sbert**[2] [14] (Sentence-BERT) model for all paragraphs from the various languages in the corpus, mapping them to a common vector space. The reliability of these embeddings is high as this model is the state-of-the-art solution for Indic cross-lingual similarity. Furthermore, data truncation was unnecessary as all data points are bound to be under 512 tokens by design, as mentioned earlier.

[2] `https://huggingface.co/l3cube-pune/indic-sentence-similarity-sbert`

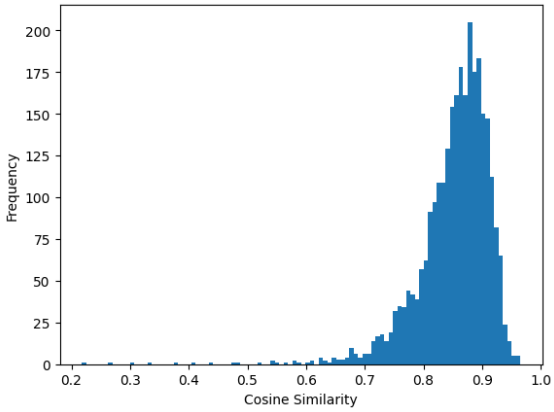


(a) en-kn

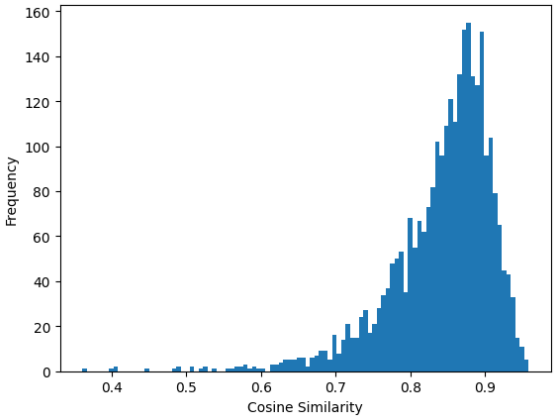
(b) en-ml



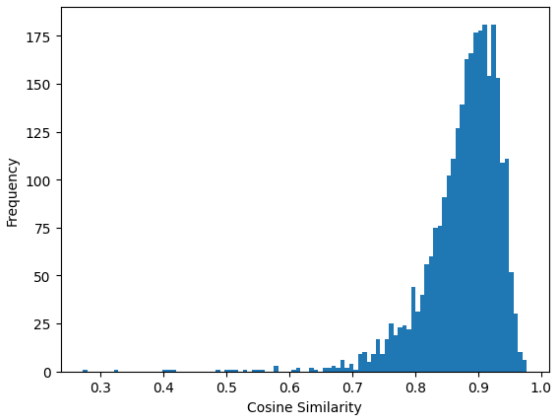
(c) en-ta



(d) en-te

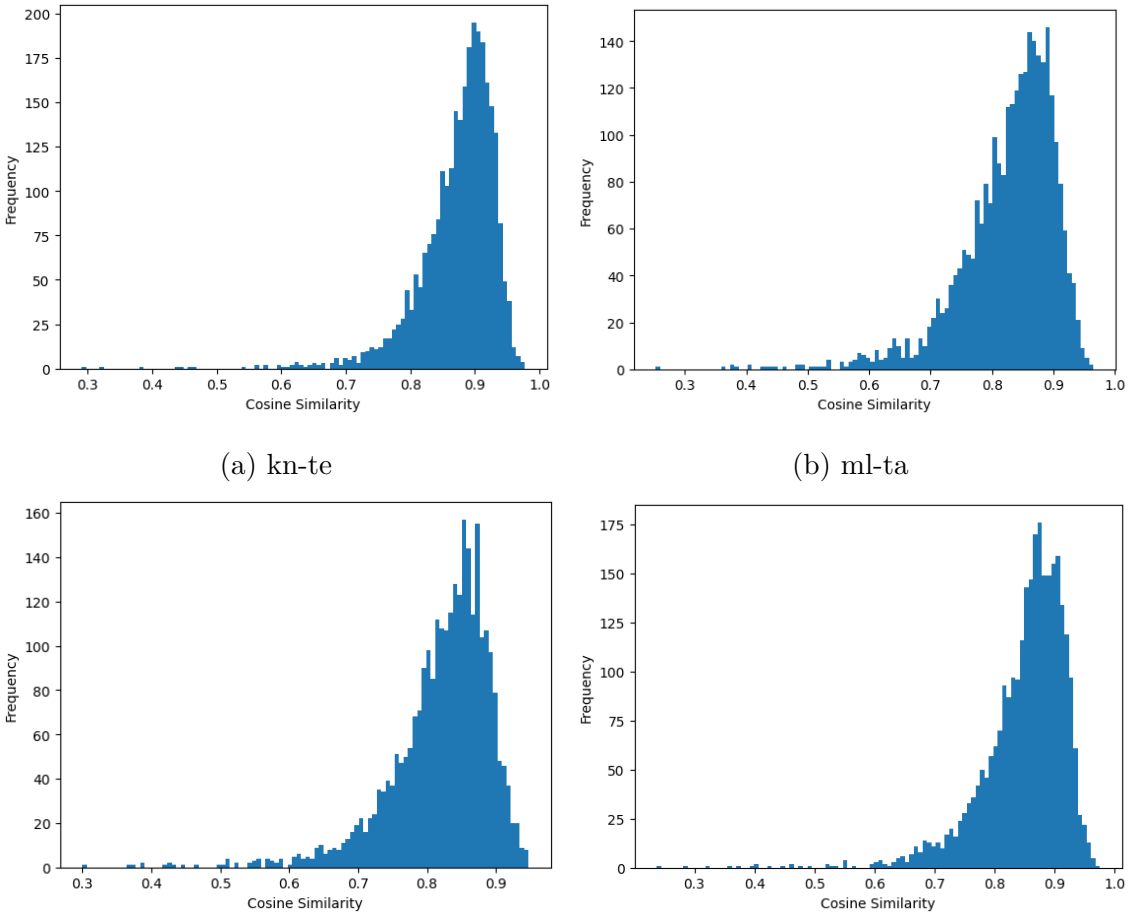


(e) kn-ml



(f) kn-ta

Figure 3.5: Distribution of alignment scores between all language pairs



(a) kn-te

(b) ml-ta

(c) ml-te

(d) ta-te

Figure 3.6: Distribution of alignment scores between all language pairs

The alignment score for the paragraphs of each possible language pair was calculated with the `cosine similarity function`. Descriptive statistics for the same are shown in Table 3.3, and the distribution of alignment scores across all tuples for all language pairs is shown in Figures 3.5 and 3.6.

$$\text{cos\_sim}(u, v) = \frac{u \cdot v}{\|u\|\|v\|} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}}$$

u and v represent the sentence embeddings, which are n-dimensional vectors containing individual component values u_i and v_i, while $|u|$ and $|v|$ represent the Euclidean norms (L2 norms) of the vectors u and v, respectively.

The examination reveals that the averages of all language tuples exceed 0.8, indicating a high degree of general similarity. In terms of the highest means and smallest standard

deviations, Kannada emerges as the subset of *CoPara* that aligns best with English. In contrast, Malayalam appears to rank lowest on these parameters.

On inspecting the tuples with lower cosine similarity (less than 0.5, which constitutes less than 3% of the data in all these XX-En tuples), it is found that some tuples contain one more or one less sentence than others. This discrepancy could be attributed to the translation process and the paragraph to which a border sentence was deemed more fitting. Another group of tuples with low performance reveals that one language in the tuple refers to a specific entity while the other merely uses a pronoun. In such cases, the context provided by embedding-based cosine similarity outside of paragraphs is limited, making it challenging to recognize that the alignment is appropriate (a limitation consistent with sentence-level aligned datasets).

3.5.2 Human evaluation

Score	English	Cross-lingual Spanish-English
5	*The two sentences are completely equivalent, as they mean the same thing.*	
	The bird is bathing in the sink. Birdie is washing itself in the water basin.	El pájaro se esta bañando en el lavabo. Birdie is washing itself in the water basin.
4	*The two sentences are mostly equivalent, but some unimportant details differ.*	
	In May 2010, the troops attempted to invade Kabul. The US army invaded Kabul on May 7th last year, 2010.	En mayo de 2010, las tropas intentaron invadir Kabul. The US army invaded Kabul on May 7th last year, 2010.
3	*The two sentences are roughly equivalent, but some important information differs/missing.*	
	John said he is considered a witness but not a suspect. “He is not a suspect anymore.” John said.	John dijo que él es considerado como testigo, y no como sospechoso. “He is not a suspect anymore.” John said.
2	*The two sentences are not equivalent, but share some details.*	
	They flew out of the nest in groups. They flew into the nest together.	Ellos volaron del nido en grupos. They flew into the nest together.
1	*The two sentences are not equivalent, but are on the same topic.*	
	The woman is playing the violin. The young lady enjoys listening to the guitar.	La mujer está tocando el violín. The young lady enjoys listening to the guitar.
0	*The two sentences are completely dissimilar.*	
	John went horse back riding at dawn with a whole group of friends. Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.	Al amanecer, Juan se fue a montar a caballo con un grupo de amigos. Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.

Figure 3.7: Annotation guidelines for annotators

A human annotator was each enlisted for a subset of each pair of Indic-English languages to assess the semantic textual similarity and to assess the quality of alignment. They were instructed to utilize the Agirre [1] guidelines and scoring system, as shown in Figure 3.7. The **LightTag.io** classification annotation tool[69] was used for the annotation. This tool features an easy-to-use interface, allowing the requisite schema and dataset to be added. Labels 0-5 were added, and the added dataset was such that each data item contained an English-Indic paragraph pair. The annotators could review their annotations and modify existing annotations. After completion of the annotation task, they were exported for analysis. The annotator view can be seen in 3.8.

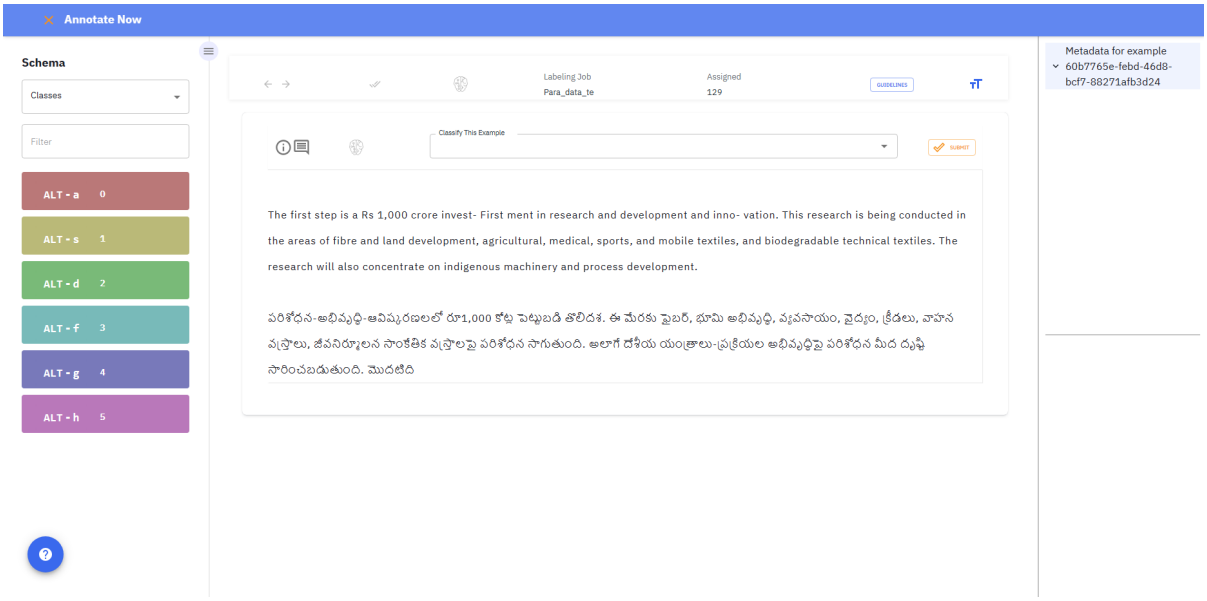


Figure 3.8: Annotator view in LightTag

There may have been discrepancies in the selection of minimum and maximum values by the annotators during the scoring process. Their scores were normalized to fall within the 0 to 1 range to accommodate these differences in scoring tendencies. This normalization process ensures a more uniform and standardized evaluation across all annotators, resulting in a more equitable appraisal of alignment quality.

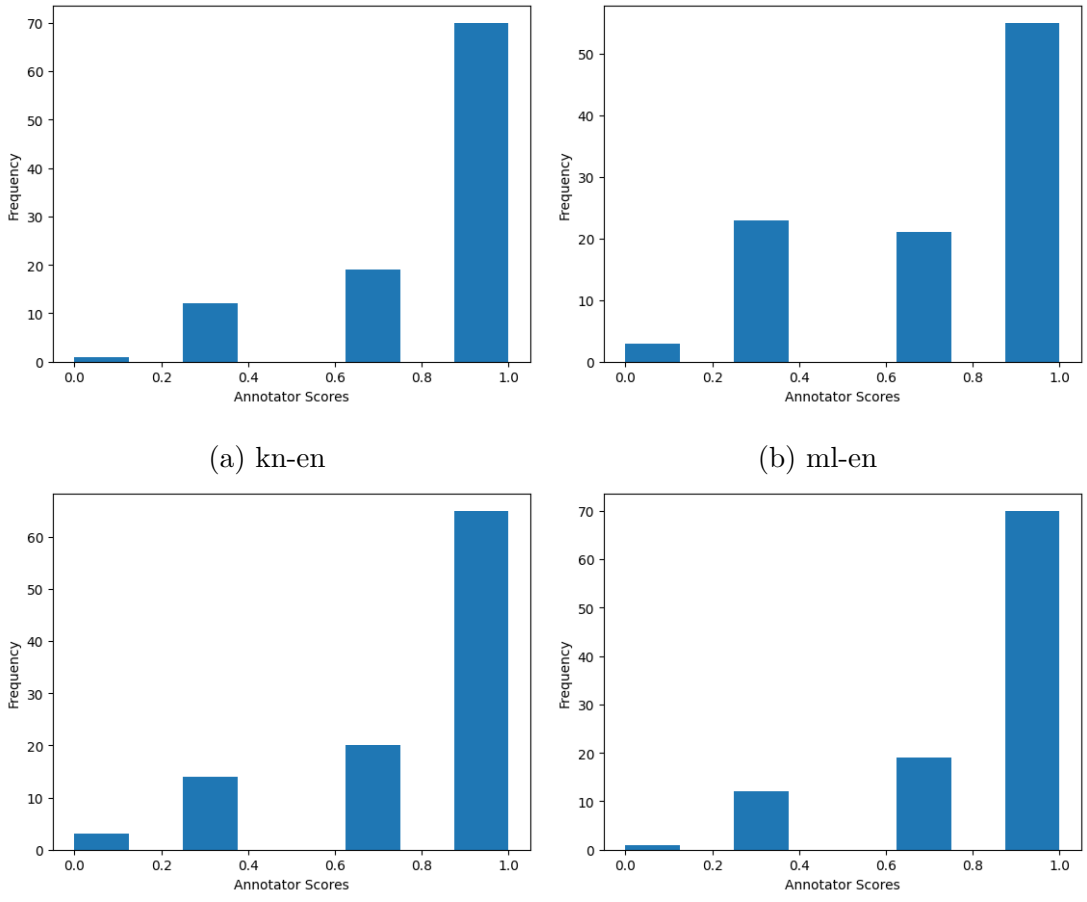
Upon human assessment, it is noticeable that the averages for all XX-En language pairs exceed 0.75, indicating a high similarity among them. Among the subsets of *CoPara*, Kannada and Telugu emerge as the most well-aligned, while Malayalam exhibits the least alignment in terms of both mean and standard deviation.

Both human and machine-generated STS scores suggest that Kannada exhibits the highest alignment with English, while Malayalam shows the lowest within the set. How-

score	**kn-en**	**ml-en**	**ta-en**	**te-en**
Mean	0.850	0.752	0.814	0.850
SD	0.245	0.301	0.278	0.245
Min	0.0	0.0	0.0	0.0
Max	1.0	1.0	1.0	1.0

Table 3.4: Descriptive statistics for xx-en human alignment scores

ever, *CoPara* as a whole is well-aligned and serves as a viable dataset for parallel paragraph-level tasks in Dravidian languages.



(a) kn-en

(b) ml-en

(c) ta-en

(d) te-en

Figure 3.9: The graphs for annotator-score (x-axis, goes from 0-1) distribution against the number of tuples (y-axis)

3.6 Sentence-Level Extension

To create sentence-level aligned datasets for each of the four Indic-English language pairs, the *l3cube-pune/indic-sentence-similarity-sbert* model was employed to generate sentence-level embeddings for all sentences within each paragraph pair. This model, known for its state-of-the-art performance in capturing cross-lingual similarity among Indic languages, provides a reliable foundation for aligning sentences across the language pairs. By fine-tuning pretrained NMT models (trained in the general domain) on the sentence-aligned extension, they can learn the intricacies and nuances of the magazine news domain, including its vocabulary, writing style, and domain-specific phrases.

Once the sentence embeddings were generated, the cosine similarity between all possible sentence pairs within each paragraph pair was calculated. This measure of similarity serves as a proxy for the alignment quality between the sentences. To ensure a high level of alignment precision, a cosine similarity threshold of 0.8 was set. Sentence pairs exhibiting a similarity score above this threshold were selected for inclusion in the new sentence-level aligned dataset. To maintain the integrity of the alignment process, all other sentence pairs containing either of the selected sentences were discarded, preventing any potential misalignments or ambiguities.

This process was repeated for all paragraph pairs across the four Dravidian-English language pairs. By leveraging the power of the SBERT model and applying a stringent cosine similarity threshold, high-quality sentence-level aligned datasets were created. The resulting sentence-level aligned datasets complement the paragraph-level corpus, providing researchers and practitioners with a comprehensive set of resources to explore and advance the field of Indic language processing. Since each paragraph can have a varying number of sentences for every language, the resulting sentence-level dataset is not n-way parallel but a collection of bilingual corpora. The unit-wise statistics for the extension are listed in Table 3.5.

Statistics	**kn**	**ml**	**ta**	**te**
Sentence Pairs	9965	7407	8167	7247
Avg. Token Length	27.0	27.9	26.2	26.3
Avg. Word Length	18.4	17.8	17.9	17.7

Table 3.5: Statistics for the sentence-level corpora created as an extension of CoPara

Chapter 4

Paragraph-level Neural Machine Translation

4.1 Introduction

Machine translation, a pivotal area within the realm of Natural Language Processing (NLP), aims to automate the process of translating text from one language to another. Traditional approaches relied on rule-based systems and statistical models, which often struggled to capture the nuances and complexities of human language. However, the advent of Neural Machine Translation (NMT) has revolutionized the field by leveraging deep learning techniques to generate more accurate and fluent translations. NMT models, typically based on recurrent or transformer architectures, learn to translate by processing entire sentences or paragraphs as holistic units, enabling them to capture context and syntactic structures more effectively. As a result, NMT has significantly improved translation quality across various language pairs and domains, paving the way for more accessible and seamless communication on a global scale. NMT utilizes artificial neural networks to generate translations by considering entire input sentences simultaneously, leading to more accurate and contextually relevant translations. Using subword embeddings has been a significant innovation in NMT, enhancing the model's ability to capture linguistic nuances and improve translation quality. However, the effectiveness of machine translation systems, including NMT, heavily relies on the availability of extensive parallel sentence corpora, which can be a limiting factor for many language pairs.

As NMT advances, researchers explore its application to Indian languages, reflecting a growing interest in leveraging advanced machine translation techniques to address the unique linguistic characteristics and challenges posed by these languages. While NMT improves translation quality for various languages, its efficacy for low-resource languages, like many Indian languages, remains a challenge. Developing multilingual NMT systems

tailored for Indian languages aims to bridge translation gaps and overcome the specific hurdles presented by low-resource languages. Recent studies have focused on enhancing multilingual NMT systems for Indic languages by incorporating state-of-the-art transformer architectures and domain-specific adaptations. Despite challenges, the research community actively explores innovative approaches to improve NMT for Indian languages, aiming to enhance translation accuracy and promote cross-lingual communication in diverse linguistic contexts.

As of the current research landscape, there is a notable absence of dedicated paragraph or passage-level NMT work tailored explicitly for Indian languages. This gap poses significant challenges in the domain of machine translation for Indian languages, as paragraph-level NMT is crucial for capturing the contextual nuances and coherence of longer text segments. The lack of paragraph-level NMT models for Indian languages can be attributed to various factors, including the scarcity of large-scale parallel corpora at the paragraph level, the complexity of linguistic structures in Indian languages, and the need for specialized domain adaptation to cater to the diverse range of topics and styles present in Indian language texts. Additionally, the computational resources required for training and fine-tuning paragraph-level NMT models for Indian languages present a substantial hurdle, further impeding progress in this area of research. Addressing these challenges and developing paragraph-level NMT models tailored for Indian languages is essential to enhance the quality and accuracy of translations, particularly for longer text segments where context plays a crucial role in conveying meaning accurately. The *Co-Para* dataset can be used to help with this very domain, and the various experiments undertaken to attempt the same are detailed in this section.

4.2 Pretrained Models for NMT

4.2.1 Choice of model

In recent years, a wide range of Transformer-based language models have been made publicly available, like mBERT and mBART which offer state-of-the-art performance on various natural language processing tasks across various languages. Architecturally, BART-based models like mBART25, mBART50 and mT5, are better suited than BERT for neural machine translation (NMT) due to their unique design based on two types of Transformers: the bidirectional encoder and the auto-regressive decoder. This architecture allows them to effectively handle both encoding and decoding tasks within

the same model, making it well-suited for sequence-to-sequence tasks like NMT. In contrast, BERT-based models, which are primarily designed as bidirectional encoders for pre-training, lack the auto-regressive decoder component necessary for generating translations in NMT tasks. These models are trained on monolingual corpora from multiple languages (often around 100) in a self-supervised manner using objectives like masked span reconstruction. They have been particularly beneficial for improving NMT performance in low-resource settings through cross-lingual transfer learning. However, these massively multilingual models have some limitations. They only cover a small fraction of the world's languages, their pre-training data is dominated by high-resource languages, the vocabulary representation for low-resource languages can be inadequate, and their large size makes them computationally expensive to train and deploy.

An alternative approach is to develop pre-trained models focused on a specific group of related languages. **IndicBART** is one such model that has been pre-trained specifically for Indian languages. It supports 11 Indian languages (Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu) plus English. IndicBART leverages the linguistic similarities and orthographic information shared across these languages. By representing all the languages in a single Devanagari script, it reduces the effective number of scripts and increases subword overlap in the vocabulary. This enables more efficient use of the modelling capacity and facilitates better cross-lingual transfer.

4.2.2 IndicBART architecture

The IndicBART model follows the same sequence-to-sequence transformer architecture as the multilingual BART (mBART) model but with a more compact design. It has 6 encoder and 6 decoder layers, a hidden size of 1024, a filter size of 4096, and 16 attention heads, resulting in approximately 245 million parameters. IndicBART uses a subword vocabulary of 64,000 tokens, generated using SentencePiece on the IndicCorp monolingual corpus covering 11 Indic languages and English. To leverage the orthographic similarity between Indic scripts, the model uses a script unification process during pretraining, mapping all Indic language data to a single script (Devanagari). The pretraining objective is similar to BART, reconstructing the original text from noisy input created by masking spans of tokens, allowing the model to learn meaningful representations of language that can be fine-tuned for downstream tasks. As a more compact model compared to mBART50 and mT5 (about 1/3 the size), IndicBART is well-suited for studying NMT and other NLG tasks on the CoPara dataset. Its smaller size and focus on Indian

languages allow the model capacity to be utilized more effectively for Indic languages. At the same time, it has been pre-trained on large amounts of monolingual data from the IndicCorp collection to build robust representations.

4.3 Finetuning IndicBART for NMT

4.3.1 Need for finetuning

IndicBART utilizes only monolingual data and a masked language modelling objective during training. The IndicCorp dataset it is trained with only comprises monolingual data. While pretraining on monolingual data enables the model to acquire general language representations, it does not provide exposure to parallel sentence pairs, which are essential for translation tasks. Consequently, the vanilla off-the-shelf IndicBART model is not suitable for neural machine translation (NMT) applications, especially at the paragraph level, without further fine-tuning or adaptation using parallel data. To effectively utilize IndicBART for downstream tasks like neural machine translation (NMT) on the CoPara dataset, it is essential to first finetune the model on parallel data, which would make it suitable for NMT. The process of finetuning IndicBART for NMT typically using a smaller learning rate compared to pretraining. This allows the model to make incremental adjustments to its parameters and learn the mapping between source and target languages while preserving the knowledge acquired during pre-training. The *IndicBART-XXEN* variant publicly-available on Hugging Face, which is finetuned on the PMI [33] and PIB [83] datasets for the Indic-English NMT task (for all 11 languages) with separate scripts, was experimented with, but the unavailability of finetuning in the other direction and usage of separate scripts (as opposed to only Devanagiri), necessitated other variants.

4.3.2 Yet Another Neural Machine Translation Toolkit (YANMTT)

YANMTT[12] is built on top of the HuggingFace Transformers library and uses a substantially modified version of the mBART model code. It provides simple scripts for NMT pretraining and fine-tuning. It enables multilingual training of sequence-to-sequence models and supports full or mixed-precision training on single or multiple GPUs. It allows finetuning of pretrained sequence-to-sequence models like mBART, BART, and

IndicBART, while providing fine-grained control over which parts of the pretrained models are used for partial initialization when finetuning. It supports lightweight fine-tuning approaches like adapter tuning and prompt tuning to allow fine-tuning of a minimal number of parameters and also enables parameter sharing across transformer layers to make the models more compact. YANMTT's fine-tuning capabilities allow users to leverage the pre-trained knowledge of IndicBART while tailoring it to the domain, style and size of the target dataset. This is particularly useful when working with a specialized corpus like CoPara, as it helps IndicBART adapt to the specific characteristics and vocabulary of the parallel paragraphs, thereby improving translation performance at the paragraph level.

4.3.3 Datasets Used

4.3.3.1 Train

The models could only be finetuned in a low-resource setting due to the availability of limited computational resources, making it difficult to use very large parallel corpora like Samanantar [75] and BPCC [25]. IndicBART has already learnt a lot of general language understanding from the pretraining it was created through. As a result, even finetuning on a low-resource parallel corpus allows the model to transfer this knowledge to the specific task and domain required. Hence, the Dravidian-English pairs from the PMI subset of the WAT 2021 MultiIndicMT10 data[63] and the same pairs from the CVIT-PIB data [83] were chosen for sentence-level finetuning. These two datasets could be used in conjunction with one another because of their shared domain, something that is in common with the CoPara data as well, increasing their suitability for finetuning. To validate this commonality, the train split for the sentence-level corpus created from the CoPara dataset is also used for finetuning (CoPara-SL). The splits for CoPara-SL are decided by picking sentences present in the parent dataset's corresponding splits. For paragraph-level training, the train split of CoPara was picked.

4.3.3.2 Validation

For the intermediate sentence-level training, the development set of WAT2021 was chosen, and for the final paragraph-level training, the development split (10%) of CoPara was picked. This was used for determining early stopping checkpoints during training.

4.3.3.3 Test

For evaluation of performance at the sentence level, the test set of WAT2021 and the *devtest* subset of FLORES101 were chosen. While the former shares the same domain as any of the training sets, the latter is more suitable for evaluating a more generalized domain. For the final performance evaluation, the test split (5%) of CoPara was picked. This was utilized for qualitative and quantitative evaluation of the final generated translations. The paragraph-level test data from CoPara was divided into five sections, numbered 1 through 5, with each section representing a specific number of sentences within a data point. The first section of the evaluation set contained only single-sentence paragraphs, with an average token length of 30, slightly higher than the 25-token average found in other sentence-level Indic datasets. This was done to ensure comparability with existing models. On the other hand, section 5, the last evaluation section, included paragraphs with more than five sentences, representing longer texts, similar to the approach taken by Zhang et al.'s hierarchical paragraph-level NMT study [95]. Sections 2, 3, and 4 corresponded to the number of sentences each passage contained within that section. Lastly, the embeddings generated from section 5 were utilized to identify evaluation paragraphs that were highly similar (>0.8) to the fine-tuning data. Out of the 130 sampled data points, 28 were found to be highly similar, resulting in a final set of 102 aligned passages for statistical analysis and inference.

4.3.4 Variants created

A total of four variant models were developed using the YANMTT toolkit and evaluated on the CoPara dataset's test split. The models were finetuned using the starting models IndicBART (base model unsuitable for NMT) and IndicBART-XXEN (finetuned variant of base model), with variations in translation direction (Dravidian-English or English-Dravidian), finetuning type (bilingual or multilingual), datasets used for finetuning (CoPara, PMI+PIB, or Sentence-level CoPara), and maximum token lengths (256 or 512 depending on the dataset).

The following presents these model variants and their respective finetuning steps taken to achieve the necessary configurations:

1. **`IB-XXEN+CoPara`**: The off-the-shelf publicly available *IndicBART-XXEN* model, already finetuned for sentence-level NMT, was further finetuned on the CoPara

dataset for each Dravidian-English language pair bilingually with a max truncation length of 512 tokens and the usage of native scripts.

2. **IB+PMI_PIB**: The basic IndicBART model, trained on the masked span reconstruction objective was finetuned for sentence-level NMT. The train split of PMI+PIB dataset, translated to Devanagari for script unification, was used for multilingual training with a max truncation length of 256 tokens and the WAT2021 development set. This model is a Many-to-One unified script variant of `IndicBART-XXEN`. Additionally, training was also done in the reverse One-to-Many direction.

3. **IB+PMI_PIB+CoPara**: The model created above was finetuned for paragraph-level NMT using a script unified version of CoPara utilizing a max truncation length of 512 tokens. The following subvariants were created for both Dravidian-English and English-Dravidian directions:

   - `IB+PMI_PIB+CoPara_Bi` was bilingually trained for each language pair.
   - `IB+PMI_PIB+CoPara_FS` was multilingually trained using only languages from the same family/high relatedness, viz. Kannada-Telugu and Malayalam-Tamil.
   - `IB+PMI_PIB+CoPara_Multi` was multilingually trained using data from all Dravidian languages.

4. **IB+PMI_PIB+CoPara-SL_Bi**: This involved an alternative finetuning of the `IB +PMI+PIB` model using the sentence-level extension of CoPara for a max truncation length of 256 tokens, to make it suitable for the CoPara domain. This was done for both Dravidian-English and English-Dravidian directions.

The finetuning process was carried out concurrently on four NVIDIA GeForce RTX 2080Ti GPUs using the Fully Sharded Data Parallel (FSDP) training strategy. FSDP is a data-parallel training approach that shards the model parameters, gradients, and optimizer states across multiple GPUs, enabling efficient memory usage and faster training times. The hyperparameters utilized were dropouts of 0.1, label smoothing of 0.1, the AdamW optimizer with a maximum learning rate of 0.001, weight decay of 0.00001, and a linear learning rate warm-up and decay scheme with 16000 warm-up steps. The models were trained until convergence on the sacreBLEU scores for the respective development sets, and the best-performing checkpoints for each language pair were saved.

4.3.5 Metrics Used

To evaluate the NMT performance of the finetuned variants, several widely used metrics were employed:

4.3.5.1 BLEU - Bilingual Evaluation Understudy

BLEU[67] is a popular metric used to automatically evaluate machine translations by measuring the similarity between a machine-generated translation and one or more human-generated reference translations. For this purpose, it counts the number of overlapping n-grams. A BLEU score can range from 0 to 1, with higher scores indicating better translation quality. However, there are various implementations of BLEU that can result in different scores for the same input due to differences in preprocessing and calculation steps. In order to address this issue and ensure reproducibility, SacreBLEU[73] was introduced as a standardized version of BLEU, through tokenization. It applies standard preprocessing steps and produces comparable BLEU scores across different systems. For standardization in this experimentation suite, the built-in `mteval-v13a` tokenizer was used for the Dravidian-English direction, while IndicNLP library's[46] Indic tokenizer was used for the English-Dravidian direction.

4.3.5.2 COMET - Crosslingual Optimized Metric for Evaluation of Translation

COMET[77] is a recently proposed learnable automatic evaluation metric for machine translation. Traditional metrics like BLEU rely on surface-level n-gram matching, but COMET leverages crosslingual pretrained language models like XLM-RoBERTa to understand deeper semantic similarities between the machine translation output and human references. It is trained on human judgments of translation quality to learn to predict human-like scores. COMET has been shown to correlate better with human assessments compared to metrics like BLEU, particularly for languages that are more distant from English. By using a learned model, COMET can account for the importance and relevance of different words and phrases to the meaning. It also has the flexibility to incorporate different types of human judgments, such as adequacy, fluency, or error categories, into the scoring. For this particular experimentation suite, the **IndicCOMET**$_{MQM}$ model, created by finetuning the state-of-the-art `Unbabel/wmt22-comet-da` model on an IndicMQM dataset[79], was utilized.

4.3.5.3 BERTScore

BERTScore[93] is a metric used to evaluate machine translation automatically that leverages pretrained contextual embeddings from BERT (Bidirectional Encoder Representations from Transformers) models. Unlike traditional metrics like BLEU that rely on exact word matches, BERTScore computes the similarity between the machine-generated translation and human reference based on their contextual embeddings. It does this by calculating a cosine similarity between the token embeddings of the candidate and reference sentences, and then taking a weighted average of the top matches. Through contextual embeddings, BERTScore is able to capture semantic similarities that go beyond surface-level word matching. It has been shown to correlate better with human judgments compared to metrics like BLEU, particularly for more abstractive translations where the wording may differ but the meaning is preserved. BERTScore is also more flexible, as it can be used with different pretrained BERT models that are trained on different languages or domains so as to fit the evaluation task better. Overall, BERTScore provides a more semantically-aware evaluation of translation quality that also aligns well with human assessments.

4.3.5.4 chrF++ - character n-gram F-score

chrF++[72] is an automatic evaluation metric for machine translation that aims to balance the strengths of character-based and word-based metrics. It is an extension of the chrF metric, which calculates the F-score of character n-grams between the machine translation output and human reference. chrF++ incorporates additional factors to improve its correlation with human judgments. Firstly, it uses a higher-order n-gram model (typically from bigrams to 6-grams) to capture more contextual information at the character level. Secondly, it applies a word bigram penalty to prevent the metric from overly rewarding fluent translations that may not preserve the meaning of the source. This helps to ensure that the translation is not only fluent but also adequate. Additionally, chrF++ includes a precision component to balance the recall-oriented nature of the original chrF metric. By combining these factors, chrF++ provides a more comprehensive and balanced evaluation of translation quality that correlates well with human assessments, particularly for languages with rich morphology or different word order than English.

4.3.5.5 BlonDe - Bilingual Evaluation of Document Translation

BlonDe[37] is a promising automatic metric for evaluating document-level machine translation quality. It goes beyond traditional sentence-level metrics like BLEU by con-

sidering discourse coherence through the categorization of discourse-related spans such as entities, tense, pronouns, and discourse markers, along with n-grams. BlonDe calculates a similarity-based F1 measure of these categorized spans between the reference and machine translations, providing a comprehensive assessment of both discourse coherence and sentence-level adequacy. Notably, BlonDe demonstrates good discriminative power across comparisons of various translation systems, including SMT, sentence-level NMT, document-level NMT, and human translations. It also correlates better with human judgments at the document level compared to other metrics. BlonDe's subvariant, dBlonDe, focuses specifically on discourse phenomena by distilling document-level translation quality, making it particularly useful for analyzing the performance of document-level NMT systems. The interpretability of BlonDe and dBlonDe's category-wise scores further facilitates error analysis and understanding strengths and weaknesses of different NMT approaches in capturing discourse-level features.

Using a combination of these metrics allows a comprehensive evaluation of the experimentation undertaken. BLEU can provide a quick and standard assessment, while COMET and BERTScore offer more semantically-aware evaluations. chrF++ can complement the other metrics by capturing character-level similarities, which is particularly relevant for Indian languages with rich morphology. BLONDE, a document-level evaluation metric, can be particularly useful in assessing the discourse coherence and context preservation capabilities of the NMT systems. By considering aspects such as entity consistency, tense, pronoun usage, and discourse markers, BLONDE provides insights into the translation quality at the paragraph level. Using multiple metrics helps mitigate the limitations of individual metrics and provides a more robust evaluation of translation quality. It allows for a holistic assessment of fluency, adequacy, semantic equivalence, and morphological consistency in the generated translations.

4.4 Results and Analysis

The performance of the IndicBART-XXEN model was compared before and after bilingual finetuning on the CoPara dataset(`IB-XXEN+CoPara_Bi`), as shown in Table 4.1. IB-XXEN+CoPara_Bi exhibited better performance than the base model across all language pairs and sections. To obtain more generalized results and test the fine-tuned model's performance improvement outside of the *CoPara* corpus, it was also evaluated on the similarly sized general-domain ***FLORES101 devtest*** [30]. Table 4.1 demon-

Language Pairs →	**kn-en**		**ml-en**		**ta-en**		**te-en**	
↓ Dataset evaluated	base	*FT*	base	*FT*	base	*FT*	base	*FT*
CoPara section 1	27.99	43.47	18.75	22.20	19.73	23.94	14.85	20.86
FLORES101 devtest	11.87	15.55	12.08	15.38	12.77	15.17	15.19	17.02
CoPara section 2	23.79	40.08	18.92	31.79	21.17	28.34	16.29	23.77
CoPara section 3	22.42	39.33	12.82	27.52	19.19	30.23	12.45	23.68
CoPara section 4	23.12	41.02	8.82	27.57	12.60	25.72	8.95	25.54
CoPara section 5	18.71	35.94	10.95	21.22	8.94	25.58	10.86	21.65
CoPara averaged	24.12	40.63	15.33	26.39	17.92	26.84	13.46	22.93

Table 4.1: BLEU scores for base(`IndicBART-XXEN`) vs FT(`IB-XXEN+CoPara_Bi`) on all XX-En pairs across all sections of CoPara and FLORES101 devtest

strates that the finetuned model performed better on this benchmark as well, across all four languages.

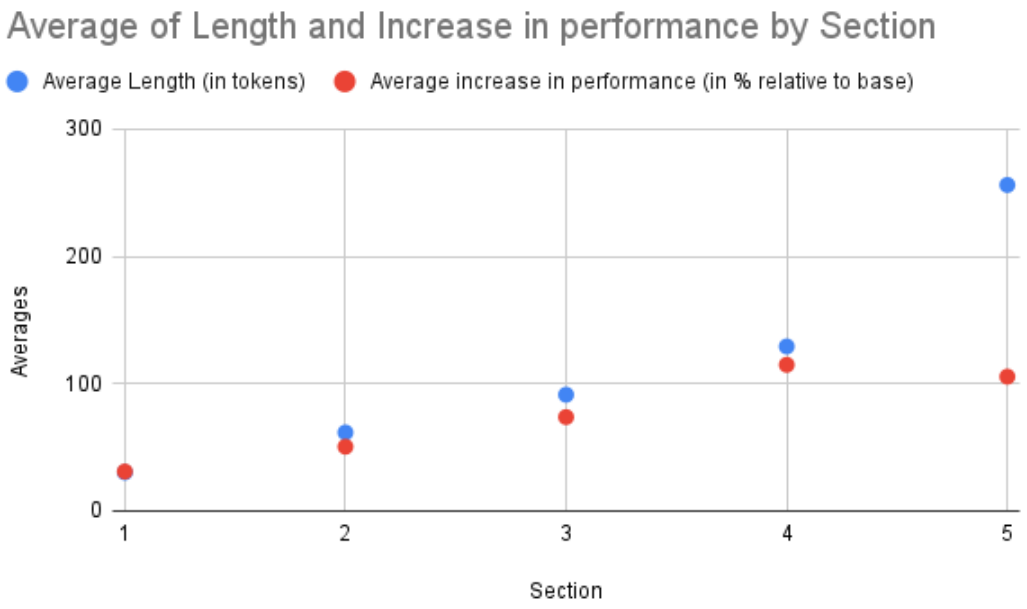


Figure 4.1: *CoPara* increases performance. Sections are representative of the number of sentences per paragraph and the Increase is in percent increase from baseline scores

Figure 4.1 provides a summary of the results from the evaluation of `IB-XXEN+CoPara_Bi` across different sections/lengths. These results indicated that the 1-sentence section of the evaluation set showed a 55% increase from the baseline BLEU scores (with similar trends for other metrics) after fine-tuning. This increase due to fine-tuning grew as the paragraphs contained more sentences, up to a composition of 4 sentences. Beyond that, there was still an improvement from the baseline, but not proportional to the increase in sentence count. This suggests that while *CoPara* enhances a model's paragraph-handling capabilities, it is limited to a certain extent, and further work would be needed for handling article-sized texts.

The effect of the standard deviation of cosine similarities for the language pairs in the CoPara dataset on the increase in scores from the baseline was independently calculated for just 1-sentence paragraphs across the 4 languages. A strong negative correlation (-0.97 Pearson's) was found between the two, indicating that as a dataset becomes less reliable in paragraph alignment, its efficiency in enriching a sentence-level NMT model decreases. This finding could explain why the Kannada fine-tuned model consistently outperformed the Malayalam one.

score	**kn-en**	**ml-en**	**ta-en**	**te-en**
Mean	0.889	0.835	0.862	0.870
SD	0.162	0.176	0.169	0.151
Min	0.4	0.4	0.4	0.4
Max	1.0	1.0	1.0	1.0

Table 4.2: Descriptive statistics for `IB-XXEN+CoPara_Bi` human evaluation scores

To further validate the effectiveness of the CoPara finetuning, human evaluation was conducted on the translations generated by the `IB-XXEN+CoPara_Bi` model, with scores ranging from 0 to 5, which were then normalized between 0 and 1 for better interpretability. The results in Table 4.2 provide further evidence of the effectiveness of finetuning on the CoPara dataset, corroborating the improvements observed in the automatic evaluation metrics. The human evaluation scores across all four Dravidian languages show high mean scores, ranging from 0.825 to 0.879, with relatively low standard deviations. Interestingly, while the automatic evaluation metrics showed some differences in performance gains across languages, with Kannada generally exhibiting the highest improvements, the human evaluation scores revealed more consistent performance across all four languages.

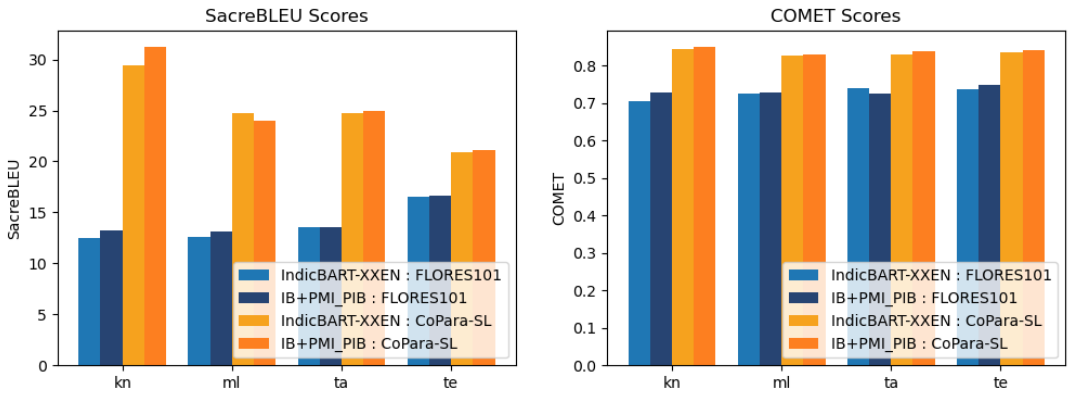


Figure 4.2: SacreBLEU and COMET metrics for bilingually(`IndicBART-XXEN`) and multilingually finetuned IndicBART(`IB+PMI_PIB`)

The effects of multilingual training were determined by comparing the ready-to-use `IndicBART-XXEN` with the multilingually finetuned `IB+PMI_PIB` by evaluating them on the FLORES101 devtest set and the CoPara-SL test set. SacreBLEU and COMET scores were calculated for Kannada, Malayalam, Tamil, and Telugu for both datasets. While the former is a surface-level n-gram matching metric, the latter captures semantic nuances and relates better to human judgments, highlighting the importance of considering multiple evaluation metrics. Multilingual finetuning with a unified script generally benefitted performance or was at least on par with the bilingual off-the-shelf model across both test sets considered, as demonstrated by Figure 4.2.

Languages→	kn-en			ml-en			ta-en			te-en		
Metric↓	*Bi*	*FS*	*Multi*	*Bi*	*FS*	*Multi*	*Bi*	*FS*	*Multi*	*Bi*	*FS*	*Multi*
SacreBLEU	45.74	46.78	44.63	28.71	28.96	27.69	29.91	30.11	27.84	24.81	25.76	23.05
chrF++	69.08	70.11	68.19	56.20	56.74	56.35	58.43	58.38	57.49	54.89	55.62	54.18
COMET	0.860	0.864	0.838	0.831	0.814	0.802	0.828	0.807	0.799	0.833	0.838	0.818
BERTScore	0.951	0.956	0.953	0.932	0.938	0.933	0.935	0.926	0.919	0.930	0.934	0.925

Table 4.3: Metrics for different finetuning strategies tested on the CoPara test set

The `IB+PMI_PIB+CoPara_Bi`, `IB+PMI_PIB+CoPara_FS` and `IB+PMI_PIB+CoPara_Multi` models were utilized to evaluate the impact of different finetuning strategies on paragraph-level NMT performance. Metric scores displayed in Table 4.3 reveal that **family-shared training slightly outperformed bilingual training**, suggesting that leveraging linguistic similarities between closely related languages can improve paragraph-level translation. However, finetuning with all four Dravidian languages generally led to a decrease

in performance compared to both bilingual and family-shared training. The performance differences between the finetuning strategies are more prominent for the SacreBLEU and chrF++ metrics compared to COMET and BERTScore. SacreBLEU and chrF++ are based on surface-level n-gram matching and character-level overlaps, respectively, while COMET and BERTScore capture more semantic-level similarities. This suggests that the choice of finetuning strategy has a bigger impact on the model's ability to generate translations that match the reference at a surface level, while the semantic-level quality is less sensitive to the finetuning approach.

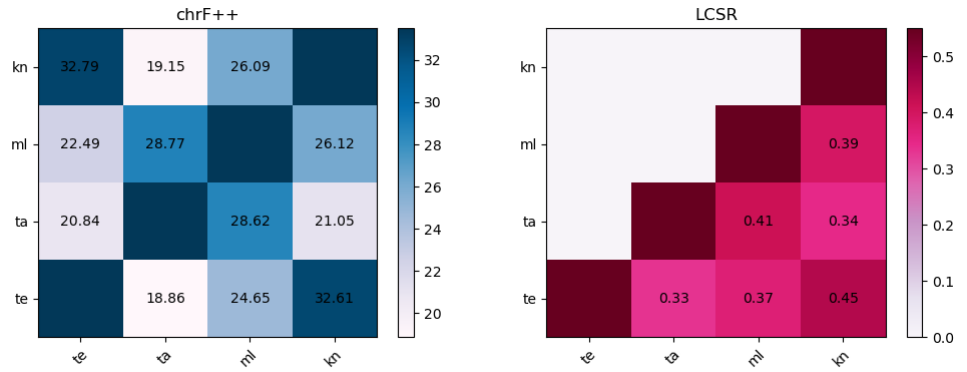


Figure 4.3: Lexical relatedness for the training sets of all language pairs

To better understand the performance differences between the different variants, the extent of lexical similarity between the train sets of the different pairs of languages was calculated using the chrF++ metric (which captures sub-word level similarities) and the LCSR (Longest Common Subsequence Ratio)[47]. Before calculation of the former, the paragraphs were tokenized so that the complex morphology of the Dravidian languages could be better handled. The scores for both metrics presented in the heatmaps in Figure 4.3 demonstrate the high degree of relatedness in the Kannada-Telugu and Tamil-Malayalam pairs. The relatedness fell off when it came to languages outside the pair, with Malayalam and Kannada exhibiting relatively high relatedness with other languages compared to Tamil and Telugu. The improved performance of Family-Sharing (FS) finetuned models over bilingually finetuned models could be linked to the high lexical relatedness amidst the language pairs chosen. Moreover, the detrimental effect of multilingual finetuning compared to bilingual finetuning could be explained by the noise induced by data from relatively distant languages being utilized. Most importantly, the magnitude of performance improvement/depreciation could be said to have a weak correlation with the lexical relatedness between the languages being added to finetuning.

Therefore, the importance of carefully selecting languages for joint training based on their relatedness is emphasized.

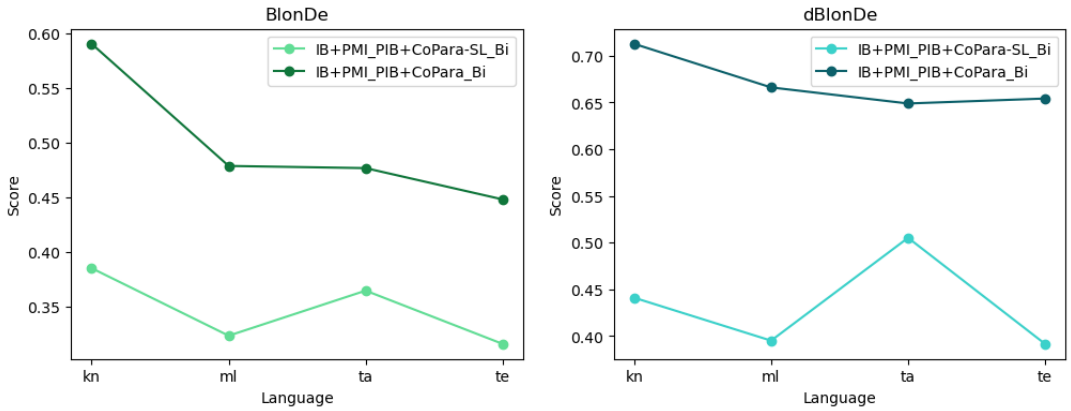


Figure 4.4: Comparison of BlonDe and dBlonDe scores for models trained at sentence-level(`IB+PMI_PIB+CoPara-SL_Bi`) and paragraph-level(`IB+PMI_PIB+CoPara`)

BlonDe combines dBlonDe, which measures discourse phenomena such as entity, tense, pronoun, and discourse markers, with a sentence-level measurement (sBlonDe) to provide a comprehensive evaluation of document-level machine translation quality. Comparing scores for the `IB+PMI_PIB+CoPara-SL_Bi` and `IB+PMI_PIB+CoPara` variants on the CoPara test set helps rule out the possibility that the model merely adapts well to the domain rather than learning to handle paragraph-level representations, as both models are exposed to the same domain through CoPara-SL and CoPara datasets, but only the latter is trained on actual paragraph-level data. The results in Figure 4.4 indicate that the **variants finetuned on the paragraph-level significantly outperform those finetuned on sentence-level across all languages**. This suggests that finetuning on the CoPara-SL dataset could improve translation performance by making the model adapt to the domain better, but it would not capture the full range of discourse-level features and coherence present in the paragraph-level CoPara dataset. These findings underscore the importance of using paragraph-level parallel data for finetuning NMT models aimed at translating longer, coherent text segments.

Finally, the performance of the `IB+PMI_PIB` and `IB+PMI_PIB+CoPara_Bi` models finetuned in the English-Dravidian direction was evaluated on the CoPara test set to assess the performance in the alternate direction. All metrics in Table 4.4 indicate that **English-Dravidian translation follows the same trend as Dravidian-English translation**, viz. variants finetuned on paragraph-level data performed significantly

better at translating paragraph-level data than those only finetuned on sentence-level data across languages. However, there was a huge disparity in the performance for the English-Dravidian direction compared to the Dravidian-English direction. This could be explained by the richer representations available for the encoder to learn in the Dravidian-English direction due to the morphologically rich nature of Dravidian languages compared to English. While the decoder generates the target translation, the quality of that output heavily depends on the representations learned by the encoder from the source data. The BlonDe metric only accepts English input for the references and hypotheses, so further analysis on the performance becomes difficult.

Language Pairs →	**en-kn**		**en-ml**		**en-ta**		**en-te**	
↓ Metric	base	*FT*	base	*FT*	base	*FT*	base	*FT*
SacreBLEU	13.37	28.23	4.93	9.16	10.11	15.66	5.50	11.56
chrF++	45.88	65.19	40.58	52.38	45.73	58.01	38.33	50.30
COMET	0.709	0.803	0.723	0.784	0.771	0.832	0.691	0.765
BERTScore	0.785	0.860	0.716	0.774	0.765	0.823	0.693	0.772

Table 4.4: Metrics for the English-Dravidian base(IB+PMI_PIB) and FT(IB+PMI_PIB+CoPara_Bi)

Chapter *5*

Leveraging Document-Level Context to Facilitate Aspect-Based Segmentation in Legal Case Files

5.1 Motivation

The Securities and Exchange Board of India, colloquially referred to as **SEBI**, is the preeminent regulatory authority entrusted with the critical oversight and governance of the securities and commodity markets within the vast expanse of the Indian subcontinent. This regulatory mandate bestowed on SEBI requires that all corporate entities operating within the purview of the securities and commodity markets must stringently adhere to and comply with the comprehensive regulations drafted and promulgated by SEBI. Adherence to these regulations requires meticulous and thorough interpretation by the companies' respective information technology departments and finance departments, as they are the primary entities tasked with ensuring regulatory compliance within their respective organizations. Furthermore, the legal representatives of these companies engage in intricate interactions and deliberations with their counterparts at SEBI, navigating the intricate web of case arguments and outcomes.

The formation of the problem statement emerged from the quest to achieve a profound comprehension of the information encapsulated within the Indian legal case files, regardless of the inherently case-specific nature of these documents. Despite the existence of numerous types of Indian legal case files and documents pertinent to this undertaking, a judicious decision was made to focus solely on a carefully curated subset of adjudication orders that specifically address insider trading violations. These adjudication orders are characterized by their convoluted and intricate language, coupled with a high degree of verbosity and subjectivity, thereby rendering any attempt at extraction or evaluation

an arduous and formidable challenge. Consequently, the broader objective evolved into identifying any sustainable patterns or trends that could potentially transcend the confines of a single set of case files, thereby enabling the extraction of relevant information ensconced within these orders or facilitating the determination of pertinent information from the textual content itself. Furthermore, it became evident that the information deemed pertinent could vary significantly from one individual to another, necessitating a tailored and customized approach. To achieve this objective, a multifaceted approach was adopted, leveraging the power of machine learning-based methods to semantically segment the legal text into ten distinct and legally applicable categories. With this sophisticated ten-way classification, the need for personalization based on an individual's level of technical knowledge or the specific requirements for a particular document can also be fulfilled.

To address the multifaceted nature of this challenge, a novel system has been meticulously developed and implemented, designed to generate separate summaries for SEBI legal case files, tailored to the specific needs and requirements of the individual requesting the summary, be they an `investor`, `an adjudicating officer`, or a `defence lawyer`.

5.2 Dataset

5.2.1 Description

The process of scraping a substantial number of adjudication orders from the Securities and Exchange Board of India (SEBI) website was undertaken. A total of *7,000* adjudication orders were scraped, out of which **27** were meticulously annotated with the assistance of a legal expert. This annotation process resulted in the creation of **2,264** sentence-label pairs, which were subsequently utilized to train various models. It should be noted that the adjudication orders chosen for annotation were specifically those pertaining to the Prohibitions of Insider Trading (PIT) regulations. This dataset is the first of its kind in the realm of Indian legal adjudication orders concerning insider trading and can be leveraged by other researchers interested in exploring this line of work.

The conceptualization of the labels was based on a preliminary study conducted on a set of case files that were randomly selected. These case files underwent a thorough examination to gain an understanding of the unique classes of sentences that each case file must possess. Subsequent to this study, each label was defined. The approach adopted was to define every label as narrowly as possible, ensuring that every unique

piece of information was captured separately. The labeling process was carried out at the sentence level, wherein each sentence in the larger case file was assigned one of the labels. To ensure that the labeling was performed with the appropriate context, the annotator considered the sentences preceding and following the target sentence to determine the appropriate label. The objective was to assign to each sentence the closest and most appropriate label of the available options. The labels utilized, their frequencies, and their descriptions are presented below:

1. `Material Facts - 533`

   (a) **Definition**: Statements containing pertinent information about the case, violation, and potential penalties crucial for determining the outcome.

   (b) **Relevant factors**: Identify the narrative of the adjudication proceedings, including details of the company, key personnel, actions perceived as violations, and the subject matter.

   (c) **Identifications made**:

   - The alleged violation;
   - Persons/companies involved;
   - Factors leading to the alleged violation;
   - Changes in shareholding patterns, etc.

2. `Procedural Facts - 310`

   (a) **Definition**: Statements about the procedure followed by authorities to initiate adjudication.

   (b) **Relevant factors**: Highlight the procedural aspects followed by the Adjudicating Officer, typically similar across cases and not offering insights into findings or conclusions.

   (c) **Identifications made**:

   - Acceptance of defense submissions;
   - Issuance of summons or show cause notices;
   - Penalty payment details (mode, channel, etc.).

3. `Statutory Facts - 190`

   (a) **Definition**: Statements invoking SEBI rules, regulations, acts, and orders, either by name/number or verbatim quotes.

(b) **Relevant factors**: Correlate various adjudication orders and sentences for similar breaches. Understand the ratio decidendi (reasoning and principles applied by the Adjudicating Officer (AO)).

(c) **Identifications made**:

- Ratio Decidendi;
- Reproduction of violated sections;
- Reproduction of penalty-defining sections.

4. `Related Facts - 95`

(a) **Definition**: General statements, truisms, reemphasis of statutory facts not constituting case facts but material for determining the outcome.

(b) **Relevant factors**: Link similar case files, understand ratio decidendi, and the perspective/stand taken by courts and likely outcomes for certain actions.

(c) **Identifications made**:

- Relevant case precedents forming the basis for courts' and AOs' stands.

5. `Issues Framed - 96`

(a) **Definition**: Statements in the form of questions, principal issues of law or fact to be adjudicated upon.

(b) **Relevant factors**: Identify the crux of the matter, formulated by the AO after considering preliminary submissions from SEBI and the defendant.

(c) **Identifications made**:

- Disputed questions;
- Disputed actions and repercussions;
- Questions to determine violations;
- Questions about actions falling within specified periods, etc.

6. `Subjective Observations - 289`

(a) **Definition**: Statements based on the personal feelings or conclusions of the AO or tribunal member, based on their reading of facts and defendant claims.

(b) **Relevant factors**: Statements informing the final order, based on determining the answers to the Issues Framed. Provide insights into the reasoning and logic for granting a particular order. May identify potential biases.

(c) **Identifications made**:

- Outcome - violations by the defendant;
- Penalty;
- Answers to Issues Framed;
- Statements reflecting the AO's perspective.

7. `Defendant Claims - 562`

(a) **Definition**: Statements elaborating the defendant's stand by countering, accepting, partially countering, or partially accepting the accusations/allegations.

(b) **Relevant factors**: Highlight the defendant's stance, reasons for non-compliance, circumstances for exceptions, etc. Considered by the AO in determining Issues Framed and the final order.

(c) **Identifications made**:

- Perspective on actions taken;
- Chronology;
- Changes in shareholding patterns;
- Case precedents favouring defendant claims.

8. `Allegations - 94`

(a) **Definition**: Statements accusing individuals or entities of violating regulations based on SEBI's understanding of facts.

(b) **Relevant factors**: Determine the charge against the defendant, identified by the action or non-action leading to the apparent violation.

(c) **Identifications made**:

- Allegedly violated provisions/sections;
- Probable penal actions for a given violation.

9. `Penalties - 50`

(a) **Definition**: Statements about the monetary penalty that should or should not be imposed, including the reason based on the regulation violated.

(b) **Relevant factors**: Determine the quantum of sentence against a determined violation.

(c) **Identifications made**:

- Relevant provisions/sections enumerating penalties for violations;
- Violation category determining penalty level;
- Penalty amount.

10. `Violations - 45`

(a) **Definition**: Statements conclusively discussing the violation of a particular regulation, act, or rule.

(b) **Relevant factors**: Determine the provisions under law that are adjudicated to be violated, based on Subjective Observation.

(c) **Identifications made**:

- Relevant violated provisions/sections;
- Violation category as per applicable sections on the defendant.

5.2.2 Quality

Out of the dataset of 27 adjudication orders used for the task, a subset of 10 documents (approximately 40% of the data) was annotated again by a second legal expert for the purpose of measuring interrater reliability. This is defined as a measurement of the degree to which the different annotators agree on the annotations assigned to the sentences. The importance of this evaluation lies in the fact that it represents the degree to which the assigned labels are indicative of the actual nature of the sentences. **`Cohen's Kappa Score`** [59] was employed to quantify this measurement. For the pairs of annotation of the chosen subset, the score was calculated to be **0.6193**, which indicates substantial agreement between the annotators.

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

κ is Cohen's kappa score, p_o is the observed agreement, and p_e is the expected agreement by chance

5.2.3 Gold Standard Summaries

Gold standard summaries for each of the three personas, namely, *Investor*, *Defense Lawyer* and *Adjudicating Officer*, were created for the 27 adjudication orders with the help of legal experts to evaluate the summarization methods after sentence classification and aspect-based filtering. Thus, a total of 81 gold summaries were made available for evaluation.

5.3 Indic Adaptation of Dataset

5.3.1 Description

To enhance the diversity and linguistic coverage of the SEBI Adjudication Orders dataset, an Indic extension was created by sentence-wise machine translation of the original English Adjudication Orders into the four Dravidian languages, namely Kannada, Malayalam, Tamil and Telugu. This extension aims to make the dataset more inclusive and accessible to researchers and legal professionals who work with Indian languages. For the construction of this extension, `IndicTrans2`[2], an open-source state-of-the-art Indic NMT model, was used. This model is trained on the Bharat Parallel Corpus Collection (BPCC), which includes both existing and new data for all scheduled Indian languages, drawing from sources like the Samanantar[75] and MASSIVE[23] datasets). This model consistently outperforms all other existing open-source models across all multi-domain benchmarks, even beating commercially available systems such as Google Translate[28]. This extension significantly increases the volume of the dataset and introduces language diversity, making it a valuable resource for multilingual legal text analysis. This process results in a total of `2,264` sentence-label pairs for all four languages which goes up to `9,056` sentence-label pairs across all languages.

5.3.2 Verification of Quality

Verifying the quality of the translations in the newly created adaption is of the utmost importance to ensure its reliability, usability, and value for future research or applications. Since there are no Dravidian language reference data points, it is important to evaluate the similarity between the translations and the original dataset to verify the usability of

the former for experiments. To this end, artificial and human evaluations were performed for the translated dataset:

`Automatic Verification.` To evaluate the quality of the translations and the resulting sentence-label pairs, a sentence similarity analysis was conducted for all source and translated pairs. The first step involved generating cross-lingual embeddings for both languages using the indic-sentence-similarity-sbert (Sentence BERT) model, which maps Indic and English representations into a shared vector space. As the state-of-the-art model for Indic cross-lingual semantic textual similarity, the embeddings generated by this model were then compared using cosine similarity. Table 5.1 presents the descriptive statistics for the obtained similarity scores. The analysis demonstrates that the mean values for each language pair surpass 0.8, suggesting a substantial level of overall resemblance, deeming the translations very usable.

statistic	**kn**	**ml**	**ta**	**te**
Mean	0.841	0.804	0.825	0.829
SD	0.079	0.075	0.080	0.074
Min	0.114	0.158	0.281	0.233
Max	0.962	0.946	0.964	0.969

Table 5.1: Descriptive statistics for cosine similarity scores with source data

`Human Verification.` To assess the quality of translations, native Dravidian language speakers who are also fluent in English are selected as human annotators. A stratified subset of 50 source-translation pairs per language, with five samples from each of the ten labels, was sampled from the parent extension dataset. Each annotator is assigned a task using the LightTag annotation tool, where they are instructed to assign a score ranging from 0 to 5 (zero alignment to high quality) to each data item, which consists of a source text and its corresponding translation within the interface. Finally, the values were normalized between 0 and 1 to provide a clearer picture. According to the annotators' analysis, the majority of the translations are nearly perfect. However, a few minor issues were observed, such as occasional spelling errors, missing numbers, and word repetitions. The mean scores for all four languages surpass 0.8, indicating a high level of alignment between the source texts and their translations. The higher standard deviation of human verification compared to artificial evaluation denotes higher variability due to lesser granularity in assigned scores. Additionally, the higher mean values compared to the same indicate that the translations are of good quality even upon consideration of

subtle linguistic factors beyond semantic similarity, like grammatical correctness, stylistic choices, fluency and coherence.

score	**kn**	**ml**	**ta**	**te**
Mean	0.862	0.821	0.840	0.869
SD	0.158	0.168	0.159	0.161
Min	0.4	0.2	0.4	0.4
Max	1.0	1.0	1.0	1.0

Table 5.2: Descriptive statistics for human alignment scores with source data

5.4 Aspect-based Segmentation of Legal Case Files

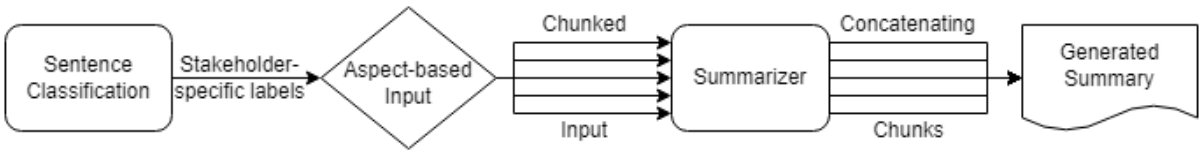


Figure 5.1: Pipeline for Aspect-based Segmentation

The following modules were used to effectively segment and summarize stakeholder-specific information from a given adjudication order (as shown in 5.1):

5.4.1 Sentence Classification Module

The legal documents are passed through the sentence classification module, which utilizes a multi-class classification model developed to classify each sentence into one of ten labels predefined in the previous section. Several techniques were explored to optimize the classifier's performance. Still, the most effective method involved fine-tuning a pre-trained BERT model on legal case files, which then leverages the contextual information from the surrounding sentences within the document to make accurate label predictions.

5.4.2 Aspect-Based Filtering Module

Different stakeholders require specific parts of the legal document based on their roles and interests, which are determined by the labels assigned in the prior step:

1. **Investors** are those who participate in the securities market, either as individuals or as organizations and are required to abide by the rules, regulations, acts, and directives established by SEBI. *Material Facts* provide them with insights into the adjudication proceedings, including information about a company's shareholdings and the alleged violations. *Penalties* are also relevant to investors, as they specify the severity of the punishment for the identified violation and the corresponding sections under various Acts that define it.

2. **Defense Lawyers** are tasked with representing their clients, whether they are individuals or companies, by engaging with their SEBI counterparts to discuss case arguments and outcomes. In addition to the *Material Facts*, *Defendant Claims* are valuable to defense lawyers as they shed light on the defendants' perspective by emphasizing their position, and *Issues Framed* help them identify the disputed questions within the case.

3. **Adjudicating Officers** of SEBI are responsible for managing case proceedings by investigating and adjudicating alleged violations. Along with *Material Facts* and *Penalties*, *Related Facts* aid them in comprehending the rationale behind the final orders and enable them to connect similar case files.

5.4.3 Summarization/Paraphrasing Module

The aspect-based summary is generated by first splitting the stakeholder-specific input, which was produced by the preceding module, into segments consisting of three sentences each. These segments are then individually processed by a state-of-the-art summarization module. Finally, the resulting outputs from the model are combined to form the comprehensive aspect-based summary that caters to the specific needs of the stakeholder in question.

5.5 Experimentation: Sentence Classification

The 27 annotated adjudication orders, along with the assigned labels were used as the dataset. A 0.75/0.10/0.15 training/validation/test split was created for the 2264 sentence-label pairs. A wide array of models, spanning from traditional machine learning algorithms to custom neural network architectures, were utilized to evaluate and enhance the performance of the sentence classification module, thereby establishing a comprehensive benchmark for the system.

5.5.1 Classical Machine Learning Methods

Different types of embeddings and traditional classification methods are employed for baseline experiments. The embeddings utilized with the Logistic Regression, Random Forest, SVM, and XGBoost methods include Word2Vec [61], tf-idf, ELMO (pretrained and the version fine-tuned on the 7,000 case files mined), and GloVe [68].

5.5.2 Classical Neural Methods

The first experiment involves the adaptation of convolutional neural networks for the classification task, drawing inspiration from the approach proposed by Yoon Kim [40], which employs trainable Word2Vec embeddings.

Subsequently, another baseline model is utilized, which incorporates GloVe embeddings for sentence representation. These embeddings are then passed through a BiLSTM and an attention mechanism to generate the predictions [97].

5.5.3 Transformer-based Methods

Various transformer models, including XLNet [89], uncased L-12 H-768 A-12 BERT (BERT-base) [16], LEGAL-BERT [6], and LEGAL-RoBERTa were experimented with, keeping each of their layers trainable and utilizing a multi-layer perceptron on top of them for label prediction. LEGAL-BERT has undergone domain adaptation using a diverse range of legal documents, such as EU legislation, UK legislation, and US court cases, among others. While using similar data, LEGAL-RoBERTa is trained using a modified strategy. The uncased L-12 H-768 A-12 BERT model was initially selected, and context was incorporated with the target sentence to two varying extents:

(i) The embeddings of the adjacent left and right sentences are concatenated with the target sentence's embedding.

(ii) A window of context (limited to a size of ten due to computational constraints) is considered at the sentence level. For the selected sentences within the window, the attention mechanism from Yang et al.'s work on Hierarchical Attention Networks [90] is employed. This mechanism assigns weights to the sentences within the window based on their importance and generates a single context vector for both the left and right context segments. This vector is then concatenated with the target sentence's embedding.

In both cases, the final input embedding is passed through a multi-layer perceptron to generate the predictions. For both these methods, the uncased L-12 H-768 A-12 model was used, which was initially fine-tuned for 70,000 steps on approximately 7,000 SEBI adjudication orders for the Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks. Finally, the context-window approach was also utilized with the LEGAL-RoBERTa model.

5.5.4 Metrics

Accuracy, macro F1 and weighted F1 scores are used to evaluate all methods used for the sentence classification module.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

TP - True Positives, TN - True Negatives, FP - False Positives, FN - False Negatives

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Macro F1} = \frac{1}{n} \sum_{i=1}^{n} \text{F1}_i$$

$$\text{Weighted F1} = \sum_{i=1}^{n} w_i \cdot \text{F1}_i$$

n is the number of classes, w_i is the proportion of samples in the i-th class, and F1_i is the F1 score for the i-th class

5.6 Experimentation: Indic Sentence Classification

Transformer-based methods have revolutionized sentence classification in Indic languages, eliminating the need for classical approaches. Multilingual models like mBERT and XLM-RoBERTa, as well as models such as IndicBERT and MuRIL capture the unique linguistic characteristics of Indic languages. By employing these transformer-based methods, state-of-the-art performance in sentence classification tasks can be achieved without the need for traditional machine learning techniques or classical neural methods. The ability of transformers to capture contextual information and long-range dependencies has proven to be highly effective in understanding and classifying sentences in Indic languages. `IndicBERT v2` [17] was the model of choice for experimentation with the Indic adaption of the SEBI dataset as it demonstrates superior performance on various benchmark datasets for Indic language tasks, including sentence classification. As a result of the increased data volume made possible due to the translation of the dataset, multilingual training can be leveraged to improve performance over monolingual classification models. The following experimentation was undertaken using pretrained transformer models:

5.6.1 Dataset, Preprocessing and Hyperparameters

Train, test and development splits created for the English sentence-label pairs were used for the Indic translation data points as well, so that training data could be used in multilingual settings without the risk of overlapping the test/development data. The scripts for the Indic data were kept untouched, because the `IndicBERTv2-MLM-Sam-TLM` variant is used, which is pretrained using separate script data. Garbage symbols at the extreme ends of the sentences are removed, extra spaces are omitted, and symbol repetition due to errors in translation is dealt with. As for the hyperparameters, the AdamW optimizer was utilized with an initial learning rate of 1e-05, and early stopping patience of two for 10 total epochs was implemented, with the finetuning being done on a 40GB NVidia A100 GPU. The batch size was modified for each method, depending on whether context was used.

5.6.2 Methods Used

1. Monolingual finetuning was undertaken for the chosen BERT model using the BertForSequenceClassification module. In addition, bilingual finetuning was un-

dertaken using the same for the Kannada-Telugu and Malayalam-Tamil language pairs. For each pair, the combination of either language's training split was used, but two runs of finetuning were undertaken using either language's development set.

2. Monolingual and multilingual finetuning was done using the chosen BERT model with the same training and validation sets as the previous step. However, the architecture is redefined to leverage document-level context using Yang et al.'s *AttentionWithContext* layer like in the previous section, which helps in capturing and aggregating the important information from the sequences at different levels of the model with different inputs. The model architecture is shown in Figure 5.2. The context window size was chosen to be ten because of computational constraints, with five sentences from either side of the target sentence being used.

5.6.3 Metrics

Like the previous section, accuracy, macro F1 and weighted F1 scores were used to assess the models finetuned using the newly created extension dataset.

5.7 Experimentation: Summarization

A meticulous evaluation of summarization models, both extractive and abstraction, is necessary for tackling the summarization of legal case files which are complex and long. To evaluate the performance of various summarization models, a set of 1,000 adjudication orders, previously labelled by the sentence classification module, was utilized as a test dataset. The model that demonstrated the highest performance on this dataset was then selected and integrated into a chunking pipeline. This pipeline was subsequently applied to a separate collection of 27 adjudication orders, for which gold standard summaries had been prepared, in order to assess its effectiveness on previously unseen data. The limited availability of legal data poses a significant challenge for training dedicated models in this domain. Consequently, the decision was made to evaluate the generalization capabilities of existing summarization models by applying them to the available dataset rather than attempting to train new models from scratch.

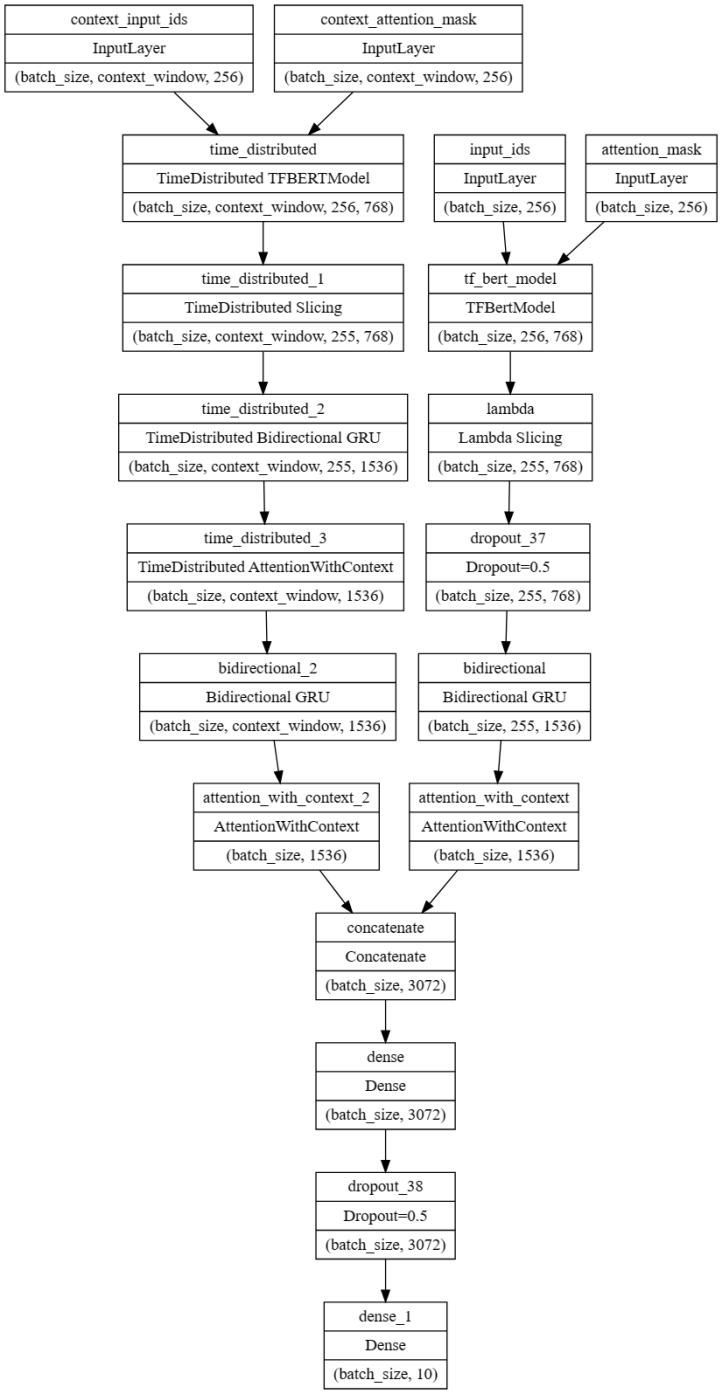


Figure 5.2: Model leveraging document-level context with output shapes

5.7.1 Unsupervised Extractive Models

Unsupervised extractive summarization models offer significant advantages for legal data, such as adaptability to the legal domain without requiring labelled data, faster processing times for lengthy legal documents, and preserving the original text, which is crucial for maintaining accuracy and transparency in legal settings. To establish a basis for comparison, several widely used unsupervised extractive summarization methods were employed, including TextRank[60], which leverages a graph-based ranking model to determine the importance of sentences based on their similarity to other sentences in the document and LexRank[21], which utilizes a similar approach but incorporates TF-IDF weights to measure sentence similarity. Additionally, the STAS(Self-Attention-based Text Summarization)[88] method was utilized, which takes advantage of self-attention mechanisms to assign saliency scores to sentences.

5.7.2 Abstractive Models

Abstractive models can generate concise and coherent summaries by understanding the context and semantics of the legal documents, allowing them to capture the key points effectively. Additionally, they can paraphrase and combine information from multiple sentences or documents, resulting in more fluent and readable summaries.

Three state-of-the-art models, namely $BERTSum_{ExtAbs}$[53], $BART$(Bidirectional and Auto-Regressive Transformers)[49] and $BRIO$(Bringing Order to Abstractive Summarization)[54], were experimented with. $BERTSum_{ExtAbs}$ utilizes a conventional encoder-decoder architecture, with BERT serving as the encoder component. On the other hand, BART is another cutting-edge summarization model that leverages a transformer-based encoder-decoder framework, which is pretrained by adding noise to its inputs and learning to reconstruct them.

Ultimately, the `BRIO` model trained on the CNN-DailyMail dataset[64] was chosen for the summarization/paraphrasing module. The model employs a non-deterministic distribution, which assigns probability mass to different candidate summaries during training based on their relative quality. Before being fed into the model, the stakeholder-specific input generated by the previous module is segmented into chunks of three sentences. The model's outputs are then concatenated to produce the final aspect-based summary. Supervised abstractive summarization models have a tendency to select information from the input based on the length of the reference summaries used during training, resulting in the generation of candidate summaries with similar lengths. However, the objective

was to retain most of the information from the input and maintain a higher recall than what is achievable with these models. A higher recall is preferable because the aim is to ensure that most sentences from the previous module are summarized in the output.

5.7.3 Metrics

The evaluation process incorporates both intrinsic and extrinsic metrics to ensure robust results. Intrinsic metrics are crucial for evaluating the quality of generated summaries without relying on human-generated reference summaries. While they offer important benefits, using extrinsic metrics as well leads to a better quality of evaluation overall.

5.7.3.1 Intrinsic

1. Semantic Similarity: This metric measures the cosine similarity between the source document and the generated summary.

2. Compression: The compression ratio is calculated by dividing the number of tokens in the summary by the number of tokens in the input document. For easier comparison, the ratio is subtracted from 1.

3. Redundancy: The average cosine similarity between all sentence pairs in a generated summary is computed to assess redundancy. Sentence embeddings are generated using the ”all-mpnet-base-v2” sentence transformer.

4. Coherence: The fluency of the generated summary is evaluated by calculating the likelihood of each sentence occurring after the previous sentence.

To enhance readability, the intrinsic scores for all personas are combined.

5.7.3.2 Extrinsic

The extrinsic metrics employed in the evaluation process include ROUGE score[50] and BERTScore[93]. ROUGE evaluates text similarity by calculating token overlaps, while BERTScore assesses similarity on a semantic level. These metrics are also presented for each of the personas.

5.8 Results and Analysis

5.8.1 Sentence Classification

Tables 5.3 and 5.4 show the performance metrics for all the neural and classical machine learning methods used for classifying sentences.

Embeddings→	Word2Vec			tf-idf			Finetuned ELMO			Pretrained ELMO			GloVe		
Classifier↓	*Acc*	$F1_m$	$F1_w$	*Acc*	$F1_m$	$F1_w$	*Acc*	$F1_m$	$F1_w$	*Acc*	$F1_m$	$F1_w$	*Acc*	$F1_m$	$F1_w$
LR	0.647	0.599	0.637	0.664	0.564	0.654	**0.706**	**0.657**	**0.699**	0.669	0.598	0.659	0.653	0.515	0.631
RF	0.660	0.622	0.660	0.671	0.671	0.671	0.656	0.595	0.648	0.596	0.505	0.596	0.578	0.550	0.577
SVM	0.675	0.607	0.665	0.679	0.613	0.669	0.648	0.593	0.635	0.608	0.491	0.599	0.653	0.555	0.643
XGBoost	0.651	0.619	0.651	0.678	0.659	0.678	0.649	0.605	0.639	0.609	0.536	0.609	0.571	0.547	0.571

Table 5.3: Classical Machine Learning Method Results

The top-performing method employs the **LEGAL-RoBERTa model in combination with a context window**, achieving an accuracy of **88.39%** and a weighted F1 score of **0.887**. Among the classical machine learning approaches, the fine-tuned ELMO embeddings demonstrate the highest performance due to their contextualized nature, deep language understanding, and incorporation of higher-level features.

Model	*Accuracy*	$F1_{macro}$	$F1_{weighted}$
LEGAL-RoBERTa + Context Window	**88.39**	**0.838**	**0.887**
Fine-tuned BERT + Context Window	83.56	0.795	0.830
Fine-tuned BERT + Two Sided Context	78.06	0.750	0.780
BERT	73.46	0.680	0.730
LEGAL-BERT	70.91	0.670	0.710
XLNet	73.29	0.620	0.700
BiLSTM + Attention	64.12	0.570	0.630
CNN-non-static	68.00	0.650	0.680

Table 5.4: Neural Method Results for Sentence Classification

Upon analyzing the performance metrics for each individual label, it became evident that the ”penalty” label yielded the lowest accuracy, while the ”procedural fact” label achieved the highest performance among all the labels considered. The inclusion of

contextual information significantly enhances the model's ability to correctly classify sentences within a document that are in close proximity to other sentences bearing the same label. This phenomenon is particularly notable for sentences labelled as "Procedural Facts" and "Subjective Observations," which tend to occur in clusters within a given document, rather than being scattered randomly throughout the text. Table 5.5 shows some examples for which the model generates correct labels.

Label	**Sentence**
material fact	The investigation covered the period from February 02, 2005 to February 14, 2005.
defendant claim	There were no such information, manipulation or use of any hidden facts to their advantage.
issue framed	Whether the Noticee is liable for imposition of monetary penalty under section 15A (b) of the SEBI Act?

Table 5.5: Examples of sentences labelled correctly by the model

5.8.2 Indic Sentence Classification

Language \ Model		BertModelForSequenceClassification			IndicBERTv2 + Context Window		
Dev/Test	Train	*Accuracy*	$F1_{macro}$	$F1_{weighted}$	*Accuracy*	$F1_{macro}$	$F1_{weighted}$
kn	kn	0.744	0.621	0.723	0.845	0.734	0.842
	kn+te	0.751	0.614	0.724	**0.891**	**0.838**	**0.889**
ml	ml	0.709	0.593	0.692	0.828	**0.784**	**0.824**
	ml+ta	0.718	0.568	0.705	**0.837**	0.765	0.815
ta	ta	0.770	0.668	0.741	0.849	0.802	0.841
	ta+ml	0.723	0.645	0.719	**0.853**	**0.804**	**0.842**
te	te	0.714	0.614	0.705	0.845	0.814	0.842
	te+kn	0.732	0.648	0.725	**0.857**	**0.829**	**0.846**

Table 5.6: Results for Dravidian Sentence Classification Methods

Based on the results obtained for the Dravidian Sentence Classification Methods shown in Table 5.5, a lot of important inferences can be made. The context-window approach consistently outperforms the baseline model for all four languages tested. Similar to the previous section, it can be inferred that the **inclusion of context enhances the model's classification performance**.

Additionally, **adding training data from a language with high linguistic similarity generally improves performance across all metrics** and both methods for all four languages tested. Kannada and Telugu, as well as Malayalam and Tamil, share many common linguistic features, such as phonology, morphology, and syntax. Due to their shared origins and geographical proximity, these language pairs have a significant amount of shared vocabulary, including cognates and loanwords. By leveraging the aforementioned shared characteristics and complementary information of closely related languages, the training data for the Kannada/Telugu and Malayalam/Tamil pairs could be augmented successfully and better performance was achieved by either method implemented. The best performing languages are Kannada and Malayalam, with weighted F1 scores of **0.889** and **0.846** respectively.

5.8.3 Summarization

Model	Semantic Similarity	Compression	Redundancy	Coherence
TextRank	0.765	0.583	0.257	0.125
LexRank	0.812	0.542	0.234	0.145
STAS	0.781	0.610	0.213	0.120
BERTSum$_{ExtAbs}$	0.776	0.878	0.193	0.327
BART	0.815	0.853	**0.154**	0.412
BRIO	**0.827**	**0.894**	0.178	**0.481**

Table 5.7: Intrinsic Metrics for Summarization

Despite their ability to identify salient information, the sentences selected by extractive methods often retain non-essential elements, a phenomenon that is evidenced by the high redundancy scores presented in Table 5.7. Furthermore, the inherent complexity and verbosity of the chosen sentences are maintained, leading to diminished coherence and minimal compression. Taken together, these findings underscore the inadequacy of extractive approaches for the specific use case under consideration, rendering them ill-suited for the task at hand.

Abstractive methods seem to outperform in generating summaries that are more succinct, unified, and easy to read. Although brevity matters, an examination of the gold standard summaries indicates that the generated output must also include all the crucial sentences from the input text, necessitating a high recall. As a result, BART and BRIO

were assessed using the chunking approach for extrinsic metrics. They exhibited strong performance, as depicted in Table 5.8, and also attained high recall. A qualitative evaluation demonstrated that the summarizer can remove non-critical expressions from the outputs while retaining the semantic content. It is apparent that abstractive techniques are better suited for this particular task.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
BART	0.416	0.261	0.352	0.792
BRIO	**0.454**	**0.271**	**0.359**	**0.809**

Table 5.8: Extrinsic Metrics for Summarization with Chunked Input

Utilizing the entire document as input without chunking led to the inclusion of arbitrary sentences unrelated to the pertinent stakeholder in the summarization process, regardless of whether extractive or abstractive methods were employed. The summarization models' lack of adaptability to the legal domain can be attributed to the fact that they are trained on datasets devoid of legal information. Consequently, aspect-based filtering using sentence classification is applied to address this limitation.

Aspect	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Adjudicating Officer	0.455	0.262	0.330	0.803
Defence Lawyer	0.424	0.241	0.320	0.798
Investor	**0.484**	**0.310**	**0.395**	**0.826**

Table 5.9: Persona-based Metrics for Chunked Input using BRIO

The persona-specific metrics in Table 5.9 indicate that the Investor summaries exhibit the best performance. This outcome is anticipated because investors require the least technical information from the adjudication order, as they only need the material facts and final penalties. In contrast, defence lawyers and adjudicating officers necessitate more intricate information, such as the issues framed and related facts. Consequently, personalized summaries can be employed to accommodate these diverse stakeholders' needs.

Chapter *6*

Conclusion and Future Work

This thesis has explored context utilization beyond the sentence level for neural machine translation and sentence classification in Dravidian languages. It has contributed to advancing Dravidian language processing by leveraging paragraph and document-level linguistic structures.

The creation of CoPara, the first publicly available paragraph-level n-way aligned corpus for the four chosen Dravidian languages and English, fills a critical gap in multilingual resources for low-resource Dravidian languages. The corpus provides a rich resource for studying cross-lingual phenomena and improving machine translation quality by offering aligned paragraphs across English and four Dravidian languages. This new dataset fills a crucial void in multilingual resources for low-resource Dravidian languages and opens up opportunities for investigating the impact of paragraph-level information on various NLP tasks, such as machine translation. Furthermore, the article-level and sentence-level extensions created using the same methodology can fuel further experimentation and contribute to the advancement of NLP research in the Dravidian language domain. The corpus can be extended to other languages in the future to enhance its diversity and enhance the possibilities of further multilingual training. Additionally, this dataset can be used for the task of Next Sentence Prediction since models can be trained to understand the logical flow of sentences across multiple languages. Through additional class-based annotation of the paragraphs in English, the dataset can be employed for Cross-Lingual Document Classification, leveraging its n-way parallelism.

Experiments conducted by fine-tuning pre-trained multilingual sequence-to-sequence models on CoPara demonstrated significant improvements in translation quality across all Dravidian-English language pairs compared to models trained using only sentence-level data. This demonstrates the utility of CoPara in leveraging paragraph-level context and multilingual training to enhance NMT systems in low-resource settings. As a result, the

significance of long-text and multilingual corpora is highlighted. The experimentation can be extended in the Indic-Indic direction because of the availability of n-way parallel data. Large Language Models pretrained on vast amounts of text data could be used to improve the accuracy and fluency of translations in Indic languages further. However, without finetuning, existing LLMs have minimal representation of Indian languages in their vocabulary and tend to hallucinate a lot with Indian languages[62]. Additionally, document-level architectures (such as hierarchical attention networks) could be incorporated to capture long-range dependencies and discourse-level information. Finally, a BlonDe-like metric could be worked on for Dravidian languages, to improve discourse-level evaluation.

Furthermore, the thesis presents an annotated dataset of SEBI legal case files with sentences classified into legally applicable categories, along with an Indic adaptation of the same. This dataset facilitated the development of a novel system for aspect-based semantic segmentation of legal case files tailored to different stakeholders. By leveraging document-level context, the system achieved improved sentence classification performance. Multilingual training using the Indic adaptation further enhanced the results, demonstrating the benefits of utilizing both document-level context and multilingual approaches. LLMs, with their extensive pretraining on diverse datasets and ability to capture long-range dependencies, could further improve the classification accuracy and generalization capabilities across different legal domains. Experimenting with larger and more advanced LLMs specifically trained on legal corpora might yield even better performance. Additionally, human evaluation of the existing sentence classification results would add to the evaluation suite. Finally, the paraphrasing module can be experimented with for Indic languages.

The experimentation and analysis conducted in this thesis underscore the significance of incorporating context beyond the sentence level and leveraging multilingual training for advancing Dravidian language processing. The datasets and methodologies introduced pave the way for future research on long-text processing and context-aware NLP tasks in low-resource languages. This thesis has made valuable contributions to the field by exploring the utilization of paragraph and document-level context for neural machine translation and sentence classification. The insights gained from this work can be extended to other low-resource languages and NLP tasks, opening up new avenues for research and development in multilingual and context-aware natural language processing.

Related Publications

1. **Nikhil E**, Mukund Choudhary, Radhika Mamidi. **CoPara: The First Dravidian Paragraph-level n-way Aligned Corpus** *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages [DravidianLangTech '23] Associated with RANLP '23*

2. **Nikhil E**, Anshul Padhi, Pulkit Parikh, Swati Kanwal, Kamal Karlapalem, Natraj Raman. **Aspect-based Summarization of Legal Case Files using Sentence Classification** *WWW '23 Companion: Companion Proceedings of the ACM Web Conference 2023*

Other Publications

1. Ishan Sanjeev Upadhyay, **Nikhil E**, Anshul Wadhawan, Radhika Mamidi. **Hopeful Men@LTEDI-EACL2021: Hope Speech Detection Using Indic Transliteration and Transformers** *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion [LTEDI '21]*

Bibliography

[1] E. Agirre, C. Banea, D. Cer, M. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, and J. Wiebe. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, and T. Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California, June 2016. Association for Computational Linguistics.

[2] AI4Bharat, J. Gala, P. A. Chitale, R. AK, S. Doddapaneni, V. Gumma, A. Kumar, J. Nawale, A. Sujatha, R. Puduppully, V. Raghavan, P. Kumar, M. M. Khapra, R. Dabre, and A. Kunchukuttan. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages, 2023.

[3] A. Bapna, I. Caswell, J. Kreutzer, O. Firat, D. van Esch, A. Siddhant, M. Niu, P. Baljekar, X. Garcia, W. Macherey, T. Breiner, V. Axelrod, J. Riesa, Y. Cao, M. X. Chen, K. Macherey, M. Krikun, P. Wang, A. Gutkin, A. Shah, Y. Huang, Z. Chen, Y. Wu, and M. Hughes. Building machine translation systems for the next thousand languages, 2022.

[4] . Borchmann, D. Wiśniewski, A. Gretkowski, I. Kosmala, D. Jurkiewicz, . Szałkiewicz, G. Pałka, K. Kaczmarek, A. Kaliska, and F. Graliński. Contract discovery: Dataset and a few-shot semantic retrieval challenge with competitive baselines. 11 2020.

[5] B. R. Chakravarthi, R. Priyadharshini, S. Banerjee, R. Saldanha, J. P. McCrae, A. K. M, P. Krishnamurthy, and M. Johnson. Findings of the shared task on machine translation in Dravidian languages. In B. R. Chakravarthi, R. Priyadharshini, A. Kumar M, P. Krishnamurthy, and E. Sherly, editors, *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 119–125, Kyiv, Apr. 2021. Association for Computational Linguistics.

[6] I. Chalkidis, I. Androutsopoulos, and N. Aletras. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy, July 2019. Association for Computational Linguistics.

[7] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos. Large-scale multi-label text classification on EU legislation. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy, July 2019. Association for Computational Linguistics.

[8] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, Nov. 2020. Association for Computational Linguistics.

[9] I. Chalkidis, N. Garneau, C. Goanta, D. M. Katz, and A. Søgaard. Lexfiles and legallama: Facilitating english multinational legal language model development, 2023.

[10] G. Choi, S. Oh, and H. Kim. Improving document-level sentiment classification using importance of sentences. *Entropy*, 22(12), 2020.

[11] R. Dabre, C. Chu, and A. Kunchukuttan. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5), sep 2020.

[12] R. Dabre, D. Kanojia, C. Sawant, and E. Sumita. YANMTT: Yet another neural machine translation toolkit. In D. Bollegala, R. Huang, and A. Ritter, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 257–263, Toronto, Canada, July 2023. Association for Computational Linguistics.

[13] R. Dabre, H. Shrotriya, A. Kunchukuttan, R. Puduppully, M. M. Khapra, and P. Kumar. Indicbart: A pre-trained model for natural language generation of indic languages. *CoRR*, abs/2109.02903, 2021.

[14] S. Deode, J. Gadre, A. Kajale, A. Joshi, and R. Joshi. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert, 2023.

[15] A. Devaraj, B. C. Wallace, I. J. Marshall, and J. J. Li. Paragraph-level Simplification of Medical Texts. *Proc Conf*, 2021:4972–4984, Jun 2021. [PubMed Central:PMC5933936] [DOI:10.18653/v1/2021.naacl-main.395] [PubMed:5302480].

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[17] S. Doddapaneni, R. Aralikatte, G. Ramesh, S. Goyal, M. M. Khapra, A. Kunchukuttan, and P. Kumar. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages, 2023.

[18] V. R. Doncel and E. M. Ponsoda. Lynx: Towards a legal knowledge graph for multilingual europe. *Law in Context. A Socio-legal Journal*, 37(1):175–178, Dec. 2020.

[19] J. Du, L. Gui, Y. He, R. Xu, and X. Wang. Convolution-based neural attention with applications to sentiment classification. *IEEE Access*, 7:27983–27992, 2019.

[20] A. El-Kishky, V. Chaudhary, F. Guzmán, and P. Koehn. CCAligned: A massive collection of cross-lingual web-document pairs. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online, Nov. 2020. Association for Computational Linguistics.

[21] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*, 22:457–479, 2004.

[22] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125, 2020.

[23] J. FitzGerald, C. Hench, C. Peris, S. Mackie, K. Rottmann, A. Sanchez, A. Nash, L. Urbach, V. Kakarala, R. Singh, S. Ranganath, L. Crist, M. Britan, W. Leeuwis, G. Tur, and P. Natarajan. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages, 2022.

[24] L. Frermann and A. Klementiev. Inducing document structure for aspect-based summarization. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th*

Annual Meeting of the Association for Computational Linguistics, pages 6263–6273, Florence, Italy, July 2019. Association for Computational Linguistics.

[25] J. Gala, P. A. Chitale, R. AK, V. Gumma, S. Doddapaneni, A. Kumar, J. Nawale, A. Sujatha, R. Puduppully, V. Raghavan, P. Kumar, M. M. Khapra, R. Dabre, and A. Kunchukuttan. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages, 2023.

[26] Z.-M. Gao and J.-A. Shih. A corpus-based computational study on translators' styles based on three chinese translations of the old man and the sea. In *The Routledge Handbook of Asian Linguistics*, pages 583–604. Routledge, 2022.

[27] R. Goebel, Y. Kano, m.-y. Kim, J. Rabelo, K. Satoh, and M. Yoshioka. Summary of the competition on legal information, extraction/entailment (coliee) 2023. pages 472–480, 09 2023.

[28] Google. Google translate. Accessed: June 10, 2023.

[29] S. Gottschalk and E. Demidova. MultiWiki. *ACM Transactions on the Web*, 11(1):1–30, feb 2017.

[30] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzman, and A. Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation, 2021.

[31] A. Gupta and K. Pala. A generic and robust algorithm for paragraph alignment and its impact on sentence alignment in parallel corpora. In *Proc of the Workshop on Indian Language Data: Resource and Evaluation(WILDRE, Organized under LREC2012)*, pages 18–27, 2012.

[32] A. Gutman and B. Avanzati. Dravidian languages, 2013.

[33] B. Haddow and F. Kirefu. Pmindia – a collection of parallel corpora of languages of india, 2020.

[34] H. Hayashi, P. Budania, P. Wang, C. Ackerson, R. Neervannan, and G. Neubig. WikiAsp: A Dataset for Multi-domain Aspect-based Summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225, 03 2021.

[35] W. Huang, J. Chen, Q. Cai, X. Liu, Y. Zhang, and X. Hu. Hierarchical hybrid neural networks with multi-head attention for document classification. *International Journal of Data Warehousing and Mining*, 18:1–16, 2022.

[36] A. J, N. Warrier, and S. Kp. Penn treebank-based syntactic parsers for south dravidian languages using a machine learning approach. *International Journal of Computer Applications*, 7, 10 2010.

[37] Y. Jiang, S. Ma, D. Zhang, J. Yang, H. Huang, and M. Zhou. Blond: An automatic evaluation metric for document-level machinetranslation. *CoRR*, abs/2103.11878, 2021.

[38] X. Kang, Y. Zhao, J. Zhang, and C. Zong. Dynamic context selection for document-level neural machine translation via reinforcement learning. *CoRR*, abs/2010.04314, 2020.

[39] A. Kapoor, M. Dhawan, A. Goel, A. T H, A. Bhatnagar, V. Agrawal, A. Agrawal, A. Bhattacharya, P. Kumaraguru, and A. Modi. HLDC: Hindi legal documents corpus. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3521–3536, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[40] Y. Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014.

[41] Y. Kim, D. T. Tran, and H. Ney. When and why is document-level context useful in neural machine translation? In A. Popescu-Belis, S. Loáiciga, C. Hardmeier, and D. Xiong, editors, *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[42] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, Sept. 13-15 2005.

[43] V. Kolipakam, F. M. Jordan, M. Dunn, S. J. Greenhill, R. R. Bouckaert, R. D. Gray, and A. Verkerk. A bayesian phylogenetic study of the dravidian language family. *Royal Society Open Science*, 5:171504, 2018.

[44] B. Krishnamurti. *The Dravidian Languages*. Cambridge Language Surveys. Cambridge University Press, 2003.

[45] S. Kudugunta, A. Bapna, I. Caswell, and O. Firat. Investigating multilingual NMT representations at scale. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th In-*

ternational Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1565–1575, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[46] A. Kunchukuttan. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`, 2020.

[47] A. Kunchukuttan and P. Bhattacharyya. Utilizing language relatedness to improve machine translation: A case study on languages of the indian subcontinent, 2020.

[48] E. Leitner, G. Rehm, and J. Moreno-Schneider. A dataset of German legal documents for named entity recognition. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France, May 2020. European Language Resources Association.

[49] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019.

[50] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[51] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[52] Y. Liu, K. Han, Z. Tan, and Y. Lei. Using context information for dialog act classification in DNN framework. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[53] Y. Liu and M. Lapata. Text summarization with pretrained encoders, 2019.

[54] Y. Liu, P. Liu, D. Radev, and G. Neubig. Brio: Bringing order to abstractive summarization, 2022.

[55] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[56] A. K. Madasamy, A. Hegde, S. Banerjee, B. R. Chakravarthi, R. Priyadharshini, H. Shashirekha, and J. McCrae. Overview of the shared task on machine translation in Dravidian languages. In B. R. Chakravarthi, R. Priyadharshini, A. K. Madasamy, P. Krishnamurthy, E. Sherly, and S. Mahesan, editors, *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 271–278, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[57] S. Mahapatra, D. Datta, S. Soni, A. Goswami, and S. Ghosh. Improving access to justice for the indian population: A benchmark for evaluating translation of legal text to indian languages, 2023.

[58] K. K. Maurya, R. Kejriwal, M. S. Desarkar, and A. Kunchukuttan. Charspan: Utilizing lexical similarity to enable zero-shot machine translation for extremely low-resource languages, 2024.

[59] M. McHugh. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82, 10 2012.

[60] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004.

[61] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality, 2013.

[62] V. Mujadia, A. Urlana, Y. Bhaskar, P. A. Pavani, K. Shravya, P. Krishnamurthy, and D. M. Sharma. Assessing translation capabilities of large language models involving english and indian languages, 2023.

[63] T. Nakazawa, H. Nakayama, I. Goto, H. Mino, C. Ding, R. Dabre, A. Kunchukuttan, S. Higashiyama, H. Manabe, W. P. Pa, S. Parida, O. Bojar, C. Chu, A. Eriguchi, K. Abe, Y. Oda, K. Sudoh, S. Kurohashi, and P. Bhattacharyya, editors. *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, Online, Aug. 2021. Association for Computational Linguistics.

[64] R. Nallapati, B. Xiang, and B. Zhou. Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023, 2016.

[65] M. Ostendorff, T. Blume, and S. Ostendorff. Towards an open platform for legal information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*,

JCDL '20, page 385–388, New York, NY, USA, 2020. Association for Computing Machinery.

[66] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[67] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[68] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

[69] T. Perry. LightTag: Text annotation platform. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 20–27, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[70] J. Philip, S. Siripragada, V. P. Namboodiri, and C. V. Jawahar. Revisiting low resource status of indian languages in machine translation. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, CODS-COMAD '21, page 178–187, New York, NY, USA, 2021. Association for Computing Machinery.

[71] P. I. B. PIB. New India Samachar, 2020.

[72] M. Popović. chrF++: words helping character n-grams. In O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, and J. Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[73] M. Post. A call for clarity in reporting BLEU scores. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, and K. Verspoor, editors,

Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics.

[74] T. Raha, S. G. Roy, U. Narayan, Z. Abid, and V. Varma. Task adaptive pretraining of transformers for hostility detection. *CoRR*, abs/2101.03382, 2021.

[75] G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. Jobanputra, R. AK, A. Sharma, S. Sahoo, H. Diddee, M. J, D. Kakwani, N. Kumar, A. Pradeep, S. Nagaraj, K. Deepak, V. Raghavan, A. Kunchukuttan, P. Kumar, and M. S. Khapra. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages, 2023.

[76] K. Ravi and V. Ravi. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46, 2015.

[77] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. COMET: A neural framework for MT evaluation. *CoRR*, abs/2009.09025, 2020.

[78] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[79] A. Sai B, T. Dixit, V. Nagarajan, A. Kunchukuttan, P. Kumar, M. M. Khapra, and R. Dabre. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada, July 2023. Association for Computational Linguistics.

[80] D. Saunders, F. Stahlberg, and B. Byrne. Using context in neural machine translation training objectives. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7764–7770, Online, July 2020. Association for Computational Linguistics.

[81] R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data. In K. Erk and N. A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

[82] R. K. Singh, M. K. Sachan, and R. B. Patel. Cross-domain sentiment classification using decoding-enhanced bidirectional encoder representations from transformers with disentangled attention. *Concurrency and Computation: Practice and Experience*, 35(6):e7589, 2023.

[83] S. Siripragada, J. Philip, V. P. Namboodiri, and C. V. Jawahar. A multilingual parallel corpora collection effort for Indian languages. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3743–3751, Marseille, France, May 2020. European Language Resources Association.

[84] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang. No language left behind: Scaling human-centered machine translation, 2022.

[85] K. Thai, M. Karpinska, K. Krishna, B. Ray, M. Inghilleri, J. Wieting, and M. Iyyer. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[86] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[87] C. Wu, F. Wu, T. Qi, and Y. Huang. Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*

Processing (Volume 2: Short Papers), pages 848–853, Online, Aug. 2021. Association for Computational Linguistics.

[88] S. Xu, X. Zhang, Y. Wu, F. Wei, and M. Zhou. Unsupervised extractive summarization by pre-training hierarchical transformers. *CoRR*, abs/2010.08242, 2020.

[89] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[90] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics.

[91] B. Zhang, A. Bapna, M. Johnson, A. Dabirmoghaddam, N. Arivazhagan, and O. Firat. Multilingual document-level translation enables zero-shot transfer from sentences to documents. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4176–4192, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[92] P. Zhang, X. Zhang, W. Chen, J. Yu, Y. Wang, and D. Xiong. Learning contextualized sentence representations for document-level neural machine translation. *CoRR*, abs/2003.13205, 2020.

[93] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert, 2020.

[94] Y. Zhang, K. Meng, and G. Liu. Paragraph-level hierarchical neural machine translation. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III*, page 328–339, Berlin, Heidelberg, 2019. Springer-Verlag.

[95] Y. Zhang, K. Meng, and G. Liu. Paragraph-level hierarchical neural machine translation. In T. Gedeon, K. W. Wong, and M. Lee, editors, *Neural Information Processing*, pages 328–339, Cham, 2019. Springer International Publishing.

[96] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*. Association for Computing Machinery, 2021.

[97] L. Zhouhan, F. Minwei, dos Santos Cicero Nogueira, Y. Mo, X. Bing, Z. Bowen, and B. Yoshua. A structured self-attentive sentence embedding. In *International Conference on Learning Representations ICLR*, 2017.

[98] J. Zhu, M. Zhu, H. Wang, and B. K. Tsou. Aspect-based sentence segmentation for sentiment summarization. In *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, TSA '09, page 65–72, New York, NY, USA, 2009. Association for Computing Machinery.