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

Nikhil E
nikhil.e@research.iiit.ac.in
IIIT Hyderabad

Mukund Choudhary
mukund.choudhary@research.iiit.ac.in
IIIT Hyderabad

Radhika Mamidi
radhika.mamidi@iiit.ac.in
IIIT Hyderabad

Abstract

We present CoPara[1], the first publicly available paragraph-level (n-way aligned) multilingual parallel corpora for Dravidian languages. The collection contains 2856 paragraph/passage pairs between English and four Dravidian languages. We source the parallel paragraphs from the New India Samachar magazine and align them with English as a pivot language. We do human and artificial evaluations to validate the high-quality alignment and richness of the parallel paragraphs of a range of lengths. To show one of the many ways this dataset can be wielded, we finetuned IndicBART, an S2S transformer model on NMT for all XX-En pairs of languages in CoPara which perform better than existing sentence-level models on standard benchmarks (like BLEU) on sentence level translations and longer text too. We show how this dataset can enrich a model trained for a task like this, with more contextual cues and beyond sentence understanding even in low-resource settings like that of Dravidian languages. Finally, the dataset and models are made available publicly at GitHub[2] to help advance research in Dravidian NLP, parallel multilingual, and beyond sentence-level tasks like NMT, etc.

1 Introduction

Public and quality Multilingual data for Indic languages, explorations in NLP for the same, and in general community interest has grown in the last few years which gave us large parallel multilingual sentence level datasets and models (Ramesh et al., 2022; AI4Bharat et al., 2023). More specifically there is also much needed and quality research coming out in Dravidian NLP through community efforts and workshops like DravidianLangTech (Madasamy et al., 2022).

However, while we improve and amass resources & techniques on Dravidian sentence-level NLP, we should also utilize publicly available data to explore if there are structures beyond sentences that can be mined to inform these growing models of longer form texts (Zhang et al., 2019). This information could lead the models one step further into solving straightforward problems like translating a paragraph (and not translating parts of it and stitching it back together) or more subtle ones like coreference resolution or author style feature distinction (Gao and Shih, 2022).

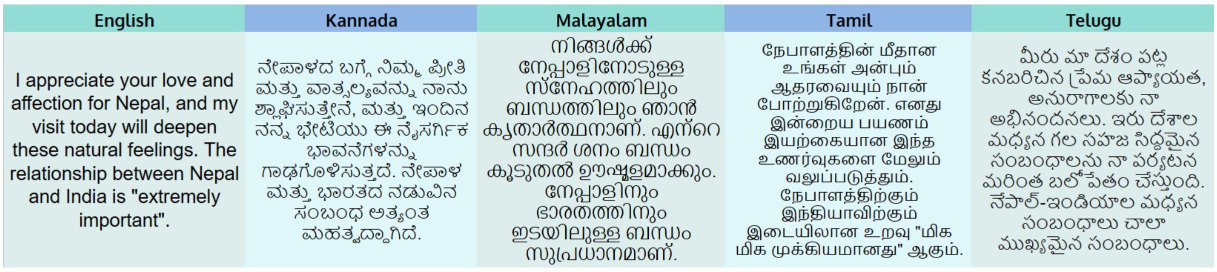
This first paragraph/passage level multilingual n-way aligned (PLA from here on) dataset in Dravidian languages thus is a step towards opening up research for these languages on document level parallel corpus creation/NMT (El-Kishky et al., 2020; Zhang et al., 2022) and be an early part of the rare but growing work in PLA corpus creation & NLP (Thai et al., 2022; Zhang et al., 2019; Devaraj et al., 2021; Gottschalk and Demidova, 2017).

The PLA work cited above and classic sentence-level aligned works (SLA, like Europarl (Koehn, 2005)) show how PLA has directly bettered NMT, has been helpful in obtaining better aligned sentence level texts, linking entities across Wikipedia like databases, better literary & medical text translation, etc. In the context of Dravidian languages, early work like (J et al., 2010) & more recently DravidianLangTech '21 (Chakravarthi et al., 2021) concluded (across 2 works) that sentence length & complexity were crucial barriers to a good translation in this language family as well. With these motivations we present *CoPara* the first public Dravidian languages' PLA corpus (example at Figure 1) which has a fair share of small & long passages, some quality checks, and show how it bettered Dravidian NMT (both sentence and paragraph level) as an example of use.

[1] *CoPara* is a reference to **copra** -meaning *the kernel of a coconut*- that lent its name to English from Dravidian languages (its a cognate - `ko[pp/bb]ar[a/i]`, in all 4 languages we explore) and the fact that the contribution is of an aligned **Para**graph **corpora**.

[2] `https://github.com/ENikhil/CoPara`

Third Workshop on Speech and Language Technologies for Dravidian Languages, pages 88–96, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_012



English	Kannada	Malayalam	Tamil	Telugu
I appreciate your love and affection for Nepal, and my visit today will deepen these natural feelings. The relationship between Nepal and India is "extremely important".	ನೇಪಾಳದ ಬಗ್ಗೆ ನಿಮ್ಮ ಪ್ರೀತಿ ಮತ್ತು ವಾತ್ಸಲ್ಯವನ್ನು ನಾನು ಶ್ಲಾಘಿಸುತ್ತೇನೆ, ಮತ್ತು ಇಂದಿನ ನನ್ನ ಭೇಟಿಯು ಈ ನೈಸರ್ಗಿಕ ಭಾವನೆಗಳನ್ನು ಗಾಢಗೊಳಿಸುತ್ತದೆ. ನೇಪಾಳ ಮತ್ತು ಭಾರತದ ನಡುವಿನ ಸಂಬಂಧ ಅತ್ಯಂತ ಮಹತ್ವದ್ದಾಗಿದೆ.	നിങ്ങൾക്ക് നേപ്പാളിനോടുള്ള സ്നേഹത്തിലും ബന്ധത്തിലും ഞാൻ കൃതാർത്ഥനാണ്. എന്റെ സന്ദർ ശനം ബന്ധം കൂടുതൽ ഊഷ്മളമാക്കും. നേപ്പാളിനും ഭാരതത്തിനും ഇടയിലുള്ള ബന്ധം സുപ്രധാനമാണ്.	நேபாளத்தின் மீதான உங்கள் அன்பும் ஆதரவையும் நான் போற்றுகிறேன். எனது இன்றைய பயணம் இயற்கையான இந்த உணர்வுகளை மேலும் வலுப்படுத்தும். நேபாளத்திற்கும் இந்தியாவிற்கும் இடையிலான உறவு "மிக மிக முக்கியமானது" ஆகும்.	మీరు మా దేశం పట్ల కనబరిచిన ప్రేమ ఆప్యాయత, అనురాగాలకు నా అభినందనలు. ఇరు దేశాల మధ్యన గల సహజ సిద్ధమైన సంబంధాలను నా పర్యటన మరింత బలోపేతం చేస్తుంది. నేపాల్-ఇండియాల మధ్యన సంబంధాలు చాలా ముఖ్యమైన సంబంధాలు.

Figure 1: An example from *CoPara*

2 Literature Review

2.1 The Dravidian Languages

Following is a brief introduction to the 4 Dravidian languages and their morpho-syntactic features of interest to motivate beyond sentence need of context in a corpus, summarised from Gutman and Avanzati (2013)'s compilation (find a more detailed sociolinguistic history at Madasamy et al. (2022)).

Some languages in the Dravidian language family (mainly South India and Asia), lack a written form, but the following 4 prominent ones have developed a large body of written literature. These agglutinative languages are mostly head-final with a flexible Subject-Object-Verb (SOV) word order, requiring the finite verb at the sentence's end. Each sentence permits one finite verb accompanied by one or more non-finite verbs. Subordinate clauses in these languages typically precede the main clause. **Kannada** is a pro-drop language that syntactically depends on case suffixes, postpositions, participles, gerunds, and infinitives. Since this makes the verb rich to express person and number, it allows subject omission. **Tamil** has its subject is usually in the nominative case. Subject-verb agreement exists, and verbs agree in person, number, and gender. It employs postpositions and case inflection to indicate syntactical relations. **Telugu**'s syntactic functions are conveyed through case suffixes and postpositions following the oblique stem. It is a pro-drop language too but lacks coordinating conjunctions, and coordinated phrases lengthen their final vowels. Relative clauses are formed using a relative participle instead of a finite verb. Finally, **Malayalam** closely resembles Tamil, but has diverged since the 8th century. It is agglutinative like Dravidian languages but has lost subject-verb and person-number agreement.

2.2 Related Work

There is some research on PLA pairs in a larger SLA dataset like Europarl but scarce work that focuses exclusively on creating a PLA corpus or NLP tasks on the same. The closest works (Thai et al., 2022; Zhang et al., 2019) respectively contribute the Par3 (multilingual, not n-way) & a Chinese-English translated novels' dataset. Thai et al. (2022) conclude that the current metrics (like BLEU (Papineni et al., 2002)) are insufficient to qualify paragraph alignment or translation etc. (thus we also do human evaluation on our data). They also find that NMT sentence-level translations are too literal as compared to human ones but BLEU prefers Google Translate over Humans. Finally they find paragraphs as a key unit for a literary paraphrase dataset and among Dravidian languages they report BLEU scores around 15-17 only for En-Ta translation quality in the literary domain. Similarly Zhang et al. (2019) innovate a hierarchical model to learn both word level and sentence level features to model a paragraph level unit in the literary domain too. They highlight how paragraph alignment is not a trivial task but make sentence alignment easier while also being a richer context for NMT models. Finally we model our Results and Analysis of the NMT experiment by plotting metric scores across sections of data with increasing average lengths.

Finally, while exploring semi-automatic methods to gauge a translator's style (Gao and Shih, 2022) highlight that sentence level auto textual aligners are more prone to errors than paragraph ones in Chinese-English and that sentence boundaries are not neatly defined by punctuation boundaries making paragraphs preferable, while (Gupta and Pala, 2012) make a Hindi-English aligner & found that it did better with aligned paragraphs, finally (Gottschalk and Demidova, 2017) showed that PLA with overlapping information in partner Wiki articles help make a comprehensive overview over shared entity facets in multilingual editions.

3 The *CoPara* Dataset

The dataset is based on the "New India Samachar" magazine (PIB, 2020), a publication launched by the Information and Broadcasting (I&B) Ministry of India in 2020 and published on a fortnightly basis in English and twelve different Indic languages. The magazine disseminates information on cabinet decisions, features content like 'Mann ki Baat', and discussions on prevailing issues.

We processed (Pipeline at Figure 2) 15 of these issues published in 2022 (subset picked randomly) to create *CoPara*. Thus we had a total of 75 issues as we aimed to align each of the 15 issues in an n-way manner across 5 languages: English and 4 Dravidian languages (Kannada, Tamil, Telugu, and Malayalam). This resulted in 2856 n-way aligned paragraphs with statistics as highlighted by Table 1. These statistics show that *CoPara* is consistent with general relative linguistic features e.g. for each sentence that is a part of an n-way aligned paragraph in English, it will be on an average longer than a Dravidian languages' sentence (more agglutinative) w.r.t. Word Length.

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 1: Average Lengths of *CoPara* paragraphs 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 too.

3.1 Data Creation

3.1.1 Imaging and Alignment

Given the characteristic presence of image-based text or non-standard encoding in all the Dravidian language magazines (mostly because they were PDFs), direct text extraction was highly erroneous. Thus the first step for an annotator (native speaker of the Dravidian language magazine is in) was to copy an English magazine content and then to capture screenshots of the corresponding Dravidian magazines for more accurate text extraction via standard Optical Character Recognition (OCR) software later.

This process involved 6 annotators who subsequently segmented these magazine contents by indicating breaks throughout the copied/screenshot magazine by using a combination of article breaks (*A*) and paragraph breaks (*#P#*). These breaks were identified by annotators after briefing them on the process of using visual cues like relative positioning, spatial heuristics, and matching design elements.

This was then checked for errors and re-annotated if required until it was satisfactorily aligned. This segmenting was served as the main way to align paragraphs across all language versions of a given magazine.

3.1.2 OCR and cleaning

For the next step of transforming these image documents into the necessary text format, we used *Google Cloud Vision*'s OCR API (Google, 2017) as it is capable of generating outputs with higher confidence than other solutions that we tested (like *Tesseract OCR* and *Amazon Textract*) and supports English & all the Dravidian languages considered.

The generated textual data was then refined by the same team of annotators to maintain quality and de-noise text-image-text conversion by tackling issues such as misinterpreted characters, incorrect order, missing words, inappropriate formatting, and other noisy artifacts.

Following this refinement, the text is subjected to further standard text pre-processing to remove extra punctuations, redundant whitespaces, line breaks, and hyperlink fixes.

3.1.3 Splitting into paragraphs

The text is then aligned using the article breaks, resulting in `1893` n-way aligned articles. To ensure accurate paragraph alignment for the next step, we cross-examine each article in all languages, tallying the number of internal paragraphs. In instances where the count is identical, the corresponding paragraphs are assumed congruent and the tuple is incorporated into the dataset (in line with assumptions of previous work) - this holds even if the count totals to a single paragraph.

Finally, we tokenize each (now aligned) paragraph from all tuples using the **IndicBART-XXEN**[3] (Dabre et al., 2021) tokenizer from the Hugging Face Library, and if the resulting token count surpasses `512`, the corresponding tuple is filtered out to allow for higher compatibility with existing language models. This rigorous procedure

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

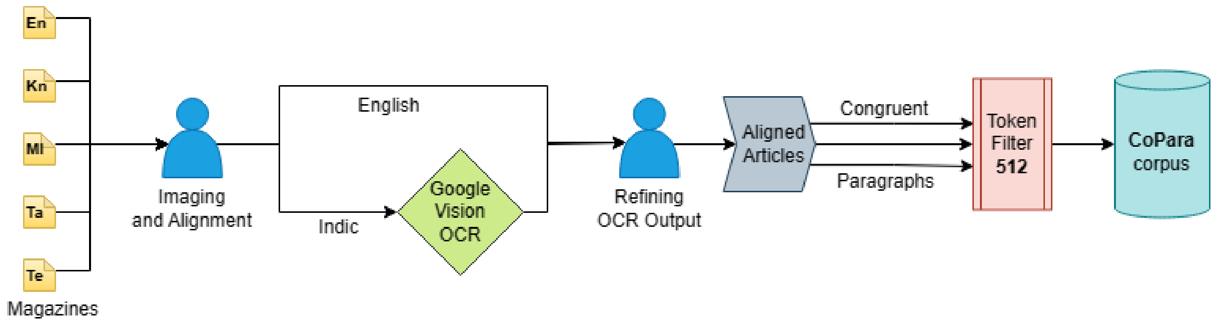


Figure 2: *CoPara* Creation Pipeline

culminates in a parallel paragraph-level corpus containing **2856** n-way aligned passages, where n = 5 (English, Kannada, Tamil, Telugu, Malayalam).

3.2 Data Quality

We adapt previous work in parallel sentences' Indic language multilingual corpus quality estimation (Ramesh et al., 2022) to fit to our paragraph-level corpus quality estimation tasks. We do two experiments to approximate the Semantic Textual Similarity (STS) of our data tuples as a proxy for their alignment quality as a correctly aligned paragraph will be generally highly similar to its other language counterpart too.

The two experiments are conducted on a randomized & length balanced subset (5%) of the data respectively an Artificial estimate (Section 3.2.1) by calculating the cosine similarity of cross-lingual embeddings and a Human estimate (Section 3.2.2) by asking human annotators to rate the same sentences on a 0-5 scale designed for an STS task.

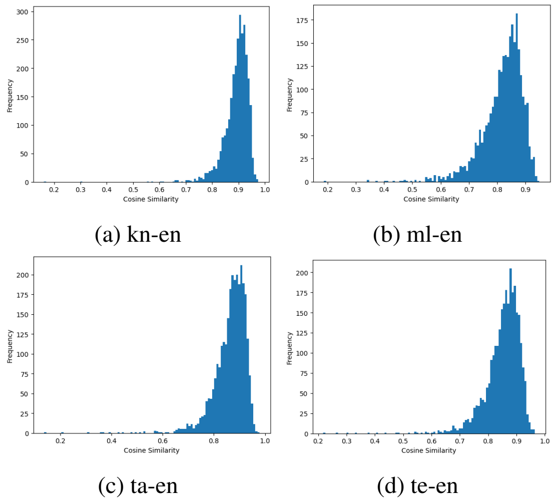
3.2.1 Artificial Alignment Evaluation

To artificially estimate how aligned our data tuples are, we generate cross-lingual embeddings using the **indic-sentence-similarity-sbert**[4] (Deode et al., 2023) (Sentence-BERT) model for all paragraphs from the different languages in our corpus to map them to a shared vector space. These embeddings are reliable as this model is the state-of-the-art for Indic cross-lingual similarity (also we did not need to truncate data as all were bound to be below 512 tokens by design).

We use these to calculate the alignment score for each **XX-En** (Dravidian-English) paragraph pair with the cosine similarity function. Descriptive statistics for the same are shown in Table 2, and the distribution of alignment scores across all tuples for all XX-En pairs are shown in Figure 3.

cos_sim	**kn-en**	**ml-en**	**ta-en**	**te-en**
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

Table 2: Descriptive statistics for xx-en alignment scores



(a) kn-en

(b) ml-en

(c) ta-en

(d) te-en

Figure 3: The graphs for cosine-similarity (x-axis, goes from 0-1) distribution against number of tuples (y-axis)

We can observe that all XX-En language tuples' averages are above 0.8 making them very similar to each other overall. Kannada comes out to be the best-aligned subset of *CoPara* in terms of the highest means and lowest standard deviations. Meanwhile, Malayalam seems the lowest in the same terms.

A closer look at the lower (<0.5, which are less than 3% data in all of these XX-En tuples) cosine-similar tuples, shows that some of these tuples have a sentence more or less than the other which would have occurred because of how the translation was

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

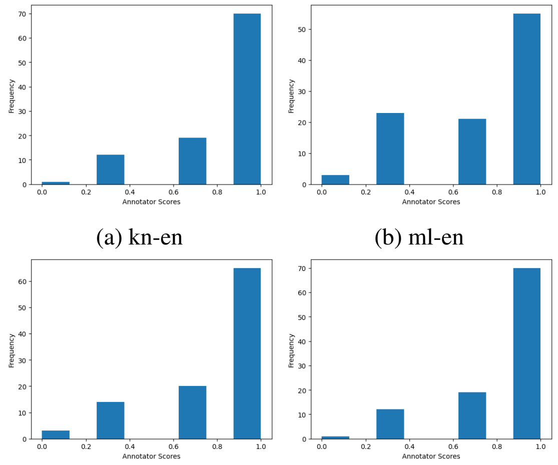
done and to which paragraph a border sentence fit more. Another set of low-performing tuples shows the existence of a referenced entity in one of the languages of the tuple and just a pronoun referral in the other. In this case, embedding-based cosine-similarity has only limited context outside of paragraphs to understand that the alignment is actually fine (this is a fallback consistent with SLA datasets).

3.2.2 Human evaluation

To assess the semantic textual similarity and gauge alignment quality, we enlisted the assistance of one human annotator for each subset of language pairs. We provided them with guidelines and utilized the scoring system from Agirre (Agirre et al., 2016), ensuring they were adequately briefed on the criteria. Annotators might have differed in their choice of minimum and maximum values when assigning scores. To account for variations in the scoring preferences of annotators, we normalized their marked scores within the range of 0 to 1. Normalizing the scores allows for a more consistent and standardized evaluation across all annotators, which would lead to a fairer assessment of the alignment quality.

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: Descriptive statistics for xx-en alignment scores



(a) kn-en

(b) ml-en

(c) ta-en

(d) te-en

Figure 4: The graphs for annotator-score (x-axis, goes from 0-1) distribution against the number of tuples (y-axis)

Based on the human evaluation, can observe that all XX-En language tuples' averages are above 0.75, making them quite similar to each other. Kannada and Telugu come out as the best-aligned subsets of *CoPara* with Malayalam being the lowest aligned in terms of the mean and the standard deviation.

We see that both human and artificial STS scores indicate that while Kannada is more highly aligned and Malayalam is the least of the set, *CoPara* as an overall set is aligned well and is a usable dataset for parallel paragraph-level tasks in Dravidian Languages. We show how it actually improves a sentence-level NMT model's performance in Dravidian languages in the next section.

4 Neural Machine Translation by Fine-tuning on *CoPara*

We now show how *CoPara* can be used to improve NMT in Dravidian languages. For the same, we take a multilingual Indic Language Model fine-tuned on the NMT task. We further finetune it on our dataset and check if the performance increases as compared to base model's results.

4.1 Dataset

As this experiment required comparable results to existing models, we sampled 5% data out from the main dataset, for evaluation. This data was a collection of five parts, labeled from 1 to 5, with each section corresponding to a specific number of sentences within a data point.

Section 1 of the evaluation set exclusively contains single sentence-length paragraphs, with an average of 30 token length (other sentence-level Indic datasets have had 25 as an average sentence length) for us to keep it comparable to existing models. In contrast, like (Zhang et al., 2019) our last evaluation section, section 5 comprises paragraphs that exceed five sentences, representing longer texts. While all the sections between 1 and 5 represent the number of sentences each passage has, in that section. Finally we used the embeddings generated from the last section to find out which evaluation paragraphs were very similar (>0.8) to fine-tuning data. We found 28 such data points out of 130 that we sampled, leaving us with 102 aligned passages to run statistics on and infer from them.

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: BLEU scores for base vs finetuned on all XX-En pairs across all sections and FLORES101 devtest

4.2 Model

For our experimentation, we employed **IndicBART**, a multilingual sequence-to-sequence Transformer model specifically pretrained on Indic languages. Architecturally, it utilizes 6 encoder and decoder layers with model and filter dimensions of 1024 and 4096 respectively on 16 attention heads (244M parameters). In specific, we utilized the publicly available ***IndicBART-XXEN*** variant, which is already finetuned on the PMI(Haddow and Kirefu, 2020) and PIB(Siripragada et al., 2020) sentence-level datasets for XX-En NMT. This will be our base model from here on.

4.3 Training

We utilized an 85:10:5 split on our dataset for training, development, and evaluation purposes. The training set was used to fine-tune the model using YANMTT(Dabre and Sumita, 2021), while the development set helped determine early stopping checkpoints. The evaluation set was utilized for qualitative, and quantitative evaluation of translations. Fine-tuning was performed individually for each XX-En language pair, using a batch size of 512 tokens for 10 epochs on an Nvidia V100 GPU. The best-performing checkpoint for each XX-En translation pair was saved as the final model.

4.4 Results

Table 4 presents a comparison of the performance of the IndicBART-XXEN model before and after fine-tuning it with our dataset. The finetuned version performs better than the base model across all language pairs and all sections. For a more generalized result and testing *CoPara* fine-tuned model's increase in performance outside of our corpus, we tested on the similarly sized ***FLORES101 devtest*** (Goyal et al., 2021) as well. Table 5 shows that the fine-tuned model does better on this benchmark as well, across the 4 languages.

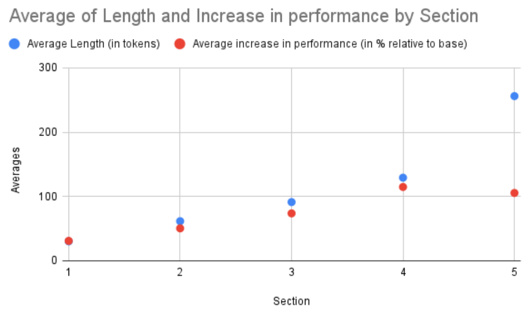


Figure 5: *CoPara* increases performance. Sections are representative of the number of sentences per paragraph and the Increase is in percent increase from baseline scores

Figure 5 summarises the results from *CoPara* evaluation, across different sections/lengths. These results show that the 1-sentence section of the evaluation set showed a 55% increase from the baseline BLEU scores (other metrics were similar) on finetuning. This increase because of finetuning increases as we make the paragraphs contain more sentences until it is composed of 4 sentences, after which there is still an increase from the baseline but not proportionate to the increase in the number of sentences. This implies that *CoPara* does increase a model's paragraph handling capabilities but only up to a certain point, more work needs to be done for article-sized texts.

Across the 4 languages, we independently also calculated the effect of the standard deviation of cosine similarities on the increase in scores from baseline on just 1-sentence paragraphs and found that there was a -0.97 (Pearson's) correlation between the two. This indicates that as a dataset gets less reliable in paragraph alignment, the efficiency of it being able to enrich a sentence-level NMT model decreases. This also helped us explain why

Language Pairs → ↓ Dataset evaluated	kn-en		ml-en		ta-en		te-en	
	base	*FT*	base	*FT*	base	*FT*	base	*FT*
CoPara section 1	0.933	0.956	0.916	0.933	0.923	0.930	0.916	0.933
FLORES101 devtest	0.933	0.925	0.925	0.932	0.927	0.925	0.920	0.936

Table 5: BERTScores for base vs finetuned on all XX-En pairs for section 1 and FLORES101 devtest

the Kannada fine-tuned model did consistently better than the Malayalam one.

Finally since BLEU has been shown insufficient as a metric for NMT (especially for longer texts), we tried out BERTScore (Zhang et al., 2020) (Table 5) to confirm results on sentence level performances of out model across languages and on teh FLORES101 devtest set as well. We see that except Kannada & Tamil data in the FLORES101 devtest sets, *CoPara* fine-tuning consistently increases performances and on these two the performance is still comparable.

5 Conclusion

We establish *CoPara* as the first parallel paragraph-level Dravidian n-way corpus by showing how it was formed and its quality by doing various evaluations. We then show that it significantly increases a sentence-level NMT model's performance on not just sentence level but also on paragraph-level data. We show that it does not hamper the base model's performance on sentence-level NMT, while enhancing it for processing paragraphs. These analyses are done on a different benchmark and on two different metrics as well.

We hope that this opens up avenues for long text and document-level NLP in Dravidian languages and that *CoPara* is grown both in quantity and quality by succeeding works.

6 Future Work and Limitations

One big improvement could be to find out better ways to align Malayalam data after finding systematic inaccuracy patterns and make it as good as Kannada data for a better *CoPara*. We can also fine-tune existing NMT models for the En-XX translation tasks and experiment with multilingual training to see if it can improve performance.

Recent work on Europarl (Amponsah-Kaakyire et al., 2021) showed that using a pivot language could cause deprecation, it would be interesting to see if the same applies to low resource language settings like Dravidian and work more on the same. (Thai et al., 2022) finds that human evaluations are still better than existing sentence-level metrics. One improvement to our work would have been to employ a more paragraph-relevant human evaluation but another improvement that is much needed is for a new set of metrics for this task.

Finally, future work can also try a hierarchical model like (Zhang et al., 2019) on our dataset to see if it can utilize the data better while we can work in parallel to make our models consume bigger paragraphs in innovative ways.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

AI4Bharat, Jay Gala, Pranjal A. Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages.

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Cristina España-Bonet, and Josef van Genabith. 2021. Do not rely on relay translations: Multilingual parallel direct Europarl. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 1–7, online. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubhanker Banerjee, Richard Saldanha, John P. McCrae, Anand Kumar M, Parameswari Krishnamurthy, and Melvin Johnson. 2021. Findings of the shared task on machine translation in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 119–125, Kyiv. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M. Khapra, and Pratyush

Kumar. 2021. Indicbart: A pre-trained model for natural language generation of indic languages. *CoRR*, abs/2109.02903.

Raj Dabre and Eiichiro Sumita. 2021. Yanmtt: Yet another neural machine translation toolkit.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert.

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

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Zhao-Ming Gao and Jou-An Shih. 2022. 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.

Google. 2017. Google cloud vision ocr api. `https://cloud.google.com/vision/docs/ocr`.

Simon Gottschalk and Elena Demidova. 2017. MultiWiki. *ACM Transactions on the Web*, 11(1):1–30.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Ankush Gupta and Kiran Pala. 2012. 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.

Alejandro Gutman and Beatriz Avanzati. 2013. [link].

Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india.

Antony J, Nandini Warrier, and Soman Kp. 2010. Penn treebank-based syntactic parsers for south dravidian languages using a machine learning approach. *International Journal of Computer Applications*, 7.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Anand Kumar Madasamy, Asha Hegde, Shubhanker Banerjee, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Hosahalli Shashirekha, and John McCrae. 2022. Overview of the shared task on machine translation in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 271–278, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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. Association for Computational Linguistics.

Press Information Bureau PIB. 2020. New India Samachar.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri, and C V Jawahar. 2020. A multilingual parallel corpora collection effort for Indian languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3743–3751, Marseille, France. European Language Resources Association.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. 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. Association for Computational Linguistics.

Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan, and Orhan Firat. 2022. Multilingual document-level translation enables zero-shot transfer from sentences to documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4176–4192, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Yuqi Zhang, Kui Meng, and Gongshen Liu. 2019. Paragraph-level hierarchical neural machine translation. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Aus-*

tralia, December 12–15, 2019, Proceedings, Part III, page 328–339, Berlin, Heidelberg. Springer-Verlag.