KT2: Kannada-Tulu Parallel Corpus Construction for Neural Machine Translation

Asha Hegde[a], **H L Shashirekha**[b]
Department of Computer Science, Mangalore University, Mangalore, Karnataka, India
{[a]ashamucs,[b]hlsrekha}@mangaloreuniversity.ac.in

Abstract

In the last decade, Neural Machine Translation (NMT) has experienced substantial advances. However, its widespread success has revealed a limitation in terms of reduced proficiency when dealing with under-resourced language pairs, mainly due to the lack of parallel corpora in comparison to high-resourced language pairs like English-German, English-Spanish, and English-French. As a result, researchers have increasingly focused on implementing NMT techniques tailored to under-resourced language pairs and thereby, the construction/collection of parallel corpora. In view of the scarcity of parallel corpus for under-resourced languages, the strategies for building a Kannada-Tulu parallel corpus and baseline models for Machine Translation (MT) of Kannada-Tulu are described in this paper. Both Kannada and Tulu languages are under-resourced due to lack of processing tools and digital resources, especially parallel corpora, which are critical for MT development. Kannada-Tulu parallel corpus is constructed in two ways: i) Manual Translation and ii) Automatic Text Generation (ATG). Various encoder-decoder based NMT approaches, including Recurrent Neural Network (RNN), Bidirectional RNN (BiRNN), and transformer-based architectures, trained with Gated Recurrent Units (GRU) and Long Short Term Memory (LSTM) units, are explored as baseline models for Kannada to Tulu (Kan-Tul) and Tulu to Kannada (Kan-Tul) sentence-level translations. Additionally, the study explores sub-word tokenization techniques for Kannada-Tulu language pairs, and the performances of these NMT models are evaluated using Character n-gram F-score (CHRF) and Bilingual Evaluation Understudy (BLEU) scores. Among the baselines, the transformer-based models outperformed other models with BLEU scores of 0.241 and 0.341 and CHRF scores of 0.502 and 0.598 for Kan-Tul and Kan-Tul sentence-level translations, respectively.

1 Introduction

Over the past two decades, Statistical Machine Translation (SMT) has held a dominant position in the field of MT (Brown et al., 1993). However, in recent years, NMT has emerged as the preferred approach, as is evident from its prevalence in majority of the shared tasks and surveys in the field of MT (Hegde et al., 2022a; Wang et al., 2021). Nevertheless, the development of efficient MT systems for under-resourced languages, which often suffer from limited resources, remains an under-explored area (Chakravarthi and Raja, 2020; Hegde et al., 2021a) highlighting the need for more attention and research in this domain (Chakravarthi et al., 2019).

In a multilingual country like India, linguistic diversity is an integral part of daily life. People use their mother tongue as the primary means of communication and employ local or regional languages as a secondary mode of communication. In Indian context, the rural population makes a significant contribution to the overall population of the country (Banerjee, 2021) and many rural residents exclusively speak their mother tongue, making translation an essential tool to bridge language barriers (Butzkamm, 2003). Translation plays a pivotal role in facilitating effective communication and fostering understanding among a culturally rich and linguistically diverse population. However, human translators may not be available everywhere and human translation is also expensive (Papineni et al., 2002). This has increased the huge demand for automatic translation viz., MT.

Kannada, the second oldest Dravidian language, holds a significant place in the linguistic landscape of India. It is predominantly spoken by the people of Karnataka, a state in southern India, where it serves as the official and administrative language. With approximately 44 million native Kannada speakers worldwide, it holds a substantial speaker

base. Over 12.6 million[1] non-Kannada speakers use Kannada as a second or third language for communication and business. Kannada has its own script derived from the Brahmi script family, consisting of 49 characters, including 13 vowels, 2 diphthongs, and 34 consonants. Linguistically, Kannada is an agglutinative language characterized by a complex morphological structure. Words in Kannada are constructed by combining compatible suffixes and/or prefixes with root words, making words meaningful. One distinctive feature of Kannada is its free word order, which allows flexibility in sentence construction. However, by convention, verbs are typically placed at the end of sentences.

Tulu, a Dravidian language, predominantly spoken by over 2.5 million people in the regions of Dakshina Kannada and Udupi in Karnataka, as well as in parts of Kasaragod in Kerala, holds a special place as the mother tongue[2] for its speakers. Tulu is distinctive among Dravidian languages for its preservation of linguistic features and unique innovations. Notably, it exhibits the use of 8 different cases (*vibhakti*), complexities in gender identification, and the common application of the ablative case, setting it apart from other Dravidian languages and highlighting its rich linguistic diversity within the Indian subcontinent. Tulu has its own script, 'Tigalari,' which is derived from the Grantha script; however, it is not widely used. Tulu's linguistic structure is characterized by a high degree of agglutination and morphology, with words formed by attaching suffixes and/or prefixes to root words, much like in Kannada. However, due to lack of Unicode support for Tigalari and influence of Kannada - the regional language, most Tulu literature and articles are written in Kannada script (Antony et al., 2016). Tulu also shares some linguistic similarities with Kannada, including verb-final inflectional patterns and relatively free word order.

MT between morphologically rich languages like Kannada and Tulu remains an uncommon and rarely explored topic (Hegde et al., 2022b). While Kannada has seen some involvement in MT, Tulu has received very limited attention in this context. However, the translation between an established local language like Kannada and an ancient regional language like Tulu, has the potential to capture the interest of MT researchers. This underscores the importance of developing efficient techniques for handling translation between these language pairs, not only for communication but also for the preservation of linguistic resources. This study also reveals the divergence patterns between Kannada and Tulu at lexical and structural levels, though both languages have overlapping vocabularies. With these objectives, this work focuses on creating a Kannada-Tulu parallel corpus for MT and setting benchmarks by implementing NMT baselines. The corpus construction is done in two ways: Manual Translation and ATG guided by linguistic rules. To benchmark the dataset, a wide range of encoder-decoder based NMT architectures, including RNN, BiRNN, and transformers with GRU and LSTM units, are implemented to translate Kannada sentences to Tulu sentences and vice versa. Further, Byte Pair Embeddings (BPE) are employed for sub-word tokenization to resolve the Out-Of-Vocabulary (OOV) issue to some extent.

The rest of the paper is organized as follows: related work is presented in Section 2 and the construction of parallel corpus is detailed in Section 3. Section 4 describes the NMT baselines followed by the experiments and results in Section 5. The paper concludes in Section 6 with avenues for future work.

2 Related Work

In NMT, a parallel corpus is essential for training the model. However, creating a high-quality parallel corpus is a challenging and time-consuming process as it requires an in-depth understanding of both source and target languages and often involves manual or semi-automatic alignment of translations, making it a critical bottleneck in building effective translation systems (Ramesh et al., 2022; Hegde and Shashirekha, 2020). Constructing a parallel corpus for under-resourced language pair introduces an additional layer of complexity due to scarcity of linguistic resources, making the process even more challenging.

The conventional method of constructing a parallel corpus relies on manual translation with the help of linguists (Hegde and Shashirekha, 2022). While this approach yields high-quality parallel data, it is expensive and laborious. Hence, MT researchers have turned their focus towards human-assisted or automatic methods for parallel corpus construc-

[1] https://censusindia.gov.in/2011Census/C-1625062018NEW.pdf

[2] https://www.mangaloretoday.com/opinion/Tulu-Language-Its\-Script-and-Dialects.html

tion, aiming to streamline the process and reduce the associated expenses (Tse et al., 2020). In this direction, web crawling (Cheok et al., 2022) and web scraping techniques (Naznin et al., 2023), are found to be more promising for constructing parallel corpora for under-resourced languages by automatically extracting bilingual content from the online resources. In addition, leveraging technologies such as the Google Translation Application Programming Interface (API) (Lowphansirikul et al., 2022), parallel text generation from speech by performing speech-to-text translation (Cettolo et al., 2012), aligned subtitles from the movies or videos (Pilevar et al., 2011) and crowd-sourced translations (Nowshin et al., 2018), aid in the construction of parallel corpora. These automated and semi-automated approaches can significantly enhance the corpus creation process. However, their effectiveness relies on post-processing for quality control to ensure accurate and reliable parallel text alignments (Soe et al., 2021).

In addition to manual and web-based methods for constructing parallel corpora, the Text Augmentation (TA) approach that comprises a dictionary-based word induction is also explored by the scholarly community (Xia et al., 2019) to enhance the size of the parallel corpus. In this method, words in an under-resourced language are replaced with similar words from a high-resource language within a sentence with the help of a bilingual dictionary. These modified sentences are subsequently refined using an unsupervised MT framework to produce augmented text data, presenting a practical solution for enriching parallel corpora to address resource issues in under-resourced languages. In addition, various TA techniques, including back-translation (Mujadia and Sharma, 2022), dictionary-based synonym replacement (Kchaou et al., 2023), word embeddings-based synonym replacement (Bayer et al., 2022), text paraphrasing (Mi et al., 2022), knowledge injection (Maharana et al., 2022), and data generation through rule-based methods (Yu et al., 2022), are also explored to enhance the size of the parallel corpora.

Though several techniques are explored to construct high-quality parallel corpora, corpus construction in morphologically rich under-resourced language pairs like Kannada-Tulu have received limited attention. In our previous work, we constructed the first-ever parallel corpus for Kannada-Tulu language pair with 30,000 parallel sentences (Hegde et al., 2022b) by adopting manual translation with the help of experts and ATG guided by named entities (considering 100 unique sentences with different combinations of person names and location). This dataset is benchmarked by implementing different NMT models (RNN, BiRNN, and transformer models with LSTM units), followed by SMT system. To the best of our knowledge, this is the only research work that has addressed the parallel corpus construction of Kannada-Tulu language pair (Hegde et al., 2022b). However, the number of parallel sentences created in this work is limited. Further, due to limited linguistic variations (just by considering the different combinations of person names and locations) the output translations are biased.

Data sources	**# of sentences**
Wikipedia	9,585
IndoWordnet	20,665
Mann ki baath	13,871
Bible	12,429
Samanantar	16,424

Table 1: Data sources and the number of Kannada sentences considered for manual translation to Tulu

3 Dataset Construction

Kannada-Tulu parallel corpus construction is carried out in two ways: Manual Translation and ATG guided by rules, and the parallel corpus construction process is given below:

3.1 Manual Translation

The process of creating Kannada-Tulu parallel corpus starts from scratch due to the limited availability of resources. Monolingual Kannada sentences are collected from digital sources and duplicate sentences are removed from the collection. The remaining sentences are then assigned to human translators for manually translating them to Tulu. Notably, Kannada script is employed for Tulu text due to the unavailability of Unicode standards for the Tulu script. The workflow for constructing the parallel corpus is shown in Figure 1 and the sources of Kannada sentences are shown in Table 1.

Manual translation, while being labor-intensive and expensive, yields a high-quality parallel corpus for MT. Existing literature highlights that human-translated sentences, despite their time and resource demands, encompass a broader range of linguistic variations and deliver precise data (Hegde

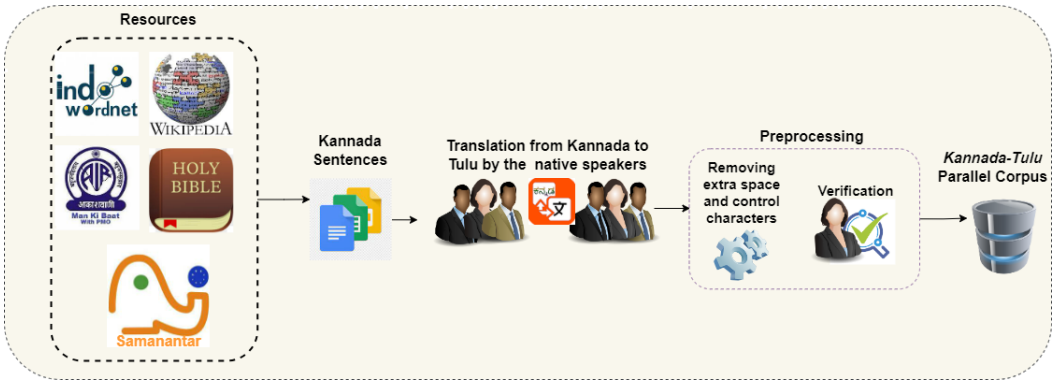

Figure 1: Framework of the Kannada-Tulu parallel corpus construction

Information of Translators		# of Translators
Gender	Male	2
	Female	10
Highest education	M.A (Tulu)	2
	M.A. (Kannada)	1
	M.Sc.	9
Medium of schooling	English	2
	Kannada	10

Table 2: Details of translators

and Shashirekha, 2022). By offering clear guidelines to translators, it is feasible to obtain a fluent and comprehensive target text, facilitating the development of efficient MT models. For manual translation, proficient native Tulu speakers who are also fluent in Kannada are chosen. Initially, they are provided with sample Kannada sentences for translation and the corresponding Tulu translations are obtained and verified. Based on the quality of the translated sentences, 12 translators are engaged for manual translation. Details of the translators are shown in Table 2 and guidelines assigned to the translators are given below:

- the target sentence must accurately reflect the context of the source sentence
- the appropriateness and fluency of the target language must be taken into account when translating
- if there are words in the source sentence that do not belong to the source language, they must be entered in the target sentence the same way as they appeared in the source sentence
- the digits must be written in the target language phonetics or in the format of the number

Kannada verb	Tulu verb	English translation		Kannada verb	Tulu verb	English translation
ಹೋಗು	ಪೋ	go		ಎಸೆ	ದಕ್ಕ್	throw
ನಿಲ್ಲು	ಉಂತು	stand		ನಗು	ತೆಲಿಪು	laugh
ನೋಡು	ತೂಲ	see		ಹೇಳು	ಪನ್	tell
ತಿನ್ನು	ತಿನ್ನ್	eat		ಬಾ	ಬಲ್ಲ	come
ಮಾತನಾಡು	ಪಾತೇರ್	speak		ಜಿಗಿ	ಲಾಗು	jump
ಓದು	ಓದು	read		ತಗೆದುಕೊಳ್ಳು	ದೊತ್ತ್	take
ಓಡು	ಬಲಿಪು	run		ಹತ್ತು	ಮಿತ್ತಾರ್	climb
ಆಡು	ಗೊಬ್ಬು	play		ಮಾಡು	ಮಲ್ವು	do
ಮಲಗು	ಜೆಪ್ಪು	sleep		ಕುಡಿ	ಪರ್	dring
ಹಿಡಿ	ಪತ್ತ್	catch		ಕೊಡು	ಕೊರು	give

Table 3: Kannada and corresponding Tulu action verbs along with their Engish translations

The resultant parallel corpus obtained from the manual translators amounts to 72,974 Kannada-Tulu sentence pairs. As Kannada sentences are collected from different resources, the parallel corpus exhibits sufficient linguistic variations.

3.2 Automatic Text Generation

Natural Language Generation (NLG) is the process of generating text based on the given input and primarily there two approaches: i) rule-based NLG and ii) template-based NLG. While rule-based NLG follows a predefined linguistic rules, template-based NLG relies on a set of predetermined templates, to generate text. Inspired by Kulkarni and Pai (2019); Hegde et al. (2022b), this work utilizes the rule-based NLG for ATG (rule-based ATG) which automatically generates the text according to the specified linguistic rules. This rule-based ATG creates new text rather than duplicating or adding slight linguistic variations to

Kannada tense suffixes	Tulu tense suffixes	Cases	Kannada	Tulu
1:['ತ್ತೇನೆ','ತ್ತಿದ್ದೇನೆ','ವೆನು','ತ್ತೇನಾ','ತ್ತಿದ್ದೇನಾ','ವೆನಾ','ವುದಿಲ್ಲ','ವೆ','ವನೇನೊ'], 2:['ತ್ತಾನೆ','ತ್ತಿದ್ದಾನೆ','ತ್ತಾನಾ','ತ್ತಿದ್ದಾನಾ','ವನು','ವನಾ','ವುದಿಲ್ಲ','ವನೇನೊ'],3: ['ತ್ತಾಳೆ','ತ್ತಿದ್ದಾಳೆ','ತ್ತಾಳಾ','ತ್ತಿದ್ದಾಳಾ','ವುದಿಲ್ಲ','ವಳು','ವಳಾ','ವಳೇನೊ'], 4:['ತ್ತಾರೆ','ತ್ತಿದ್ದಾರೆ','ತ್ತಾರಾ','ತ್ತಿದ್ದಾರಾ','ವರು','ವರಾ','ವರೇನೊ','ವುದಿಲ್ಲ']	1:['ಪೆ','ವೊಂದುಲ್ಲೆ','ಪೆ','ಪೆನಾ','ವೊಂದುಲ್ಲೆನಾ','ಪೆನಾ','ಪುಜೆ','ಪೆ','ಪೆನ ದಾನ'], 2:['ಪೆ','ವೊಂದುಲ್ಲೆ','ಪೆನಾ','ವೊಂದುಲ್ಲೆನಾ','ಪೆ','ಪೆನಾ','ಪುಜೆ','ಪೆನ ದಾನ'], 3: ['ಪಲ್','ವೊಂದುಲ್ಲಲ್','ಪಲಾ','ವೊಂದುಲ್ಲಲಾ','ಪುಜಲ್','ವಲ್','ವಲಾ','ವಲಾ ದಾನ'], 4: ['ಪೇರ್', 'ವೊಂದುಲ್ಲೇರ್', 'ಪೆರಾ', 'ವೊಂದುಲ್ಲೆರಾ' ,'ವೇರ್', 'ವೆರಾ', 'ಪೇರಾ ದಾನ', 'ಪುಜೆರ್']	Dative	['ಕ್ಕೆ','ಗೆ']	['ಕ್','ಗ್']
		Accusative	['ವನ್ನು','ಯನ್ನು']	['ನ್']
		Locative	['ದಲ್ಲಿ','ಯಲ್ಲಿ']	['ಡ್','ಟ್']
		Ablative	['ದಿಂದ','ಯಿಂದ']	['ಡ್ದ್']

Table 4: Sample tense suffixes and *vibhakti* suffixes for Kannada and Tulu langugaes

the existing data. This attempt helps to resolve the lack of resource issue to some extent in NLP tasks like MT and question and answering for under-resourced languages. Though rule-based ATG creates slightly narrow/biased data, its strength lies in the precise application of linguistic rules. Additionally, it serves as a helpful resource for language learners, offering the generation of straightforward sentences that can aid second language acquisition. Second-language learners can utilize such modules to generate sentences in a controlled manner and learn these languages. This computational tool can benefit users who know any one language and are willing to learn another language.

The proposed rule-based ATG involves generating text by randomly selecting subjects, objects and verbs from the list of subjects, objects, and verbs respectively. These subjects, objects and verbs which are used in day-to-day conversation are collected from the available online resources. Verbs take different forms when combined with the suffixes depending on the grammatical function they serve and the phenomenon is called verbal inflection. The verbal inflections are not only the tense markers but also encode Person-Gender- Number (PNG) information with respect to the subject. However, for morphologically rich languages like Kannada and Tulu, this verbal inflections become more challenging to handle as the root verb changes its spelling leading to ambiguity during categorizing the verbs and objects. The steps involved in developing the proposed rule-based ATG system are given below:

1. 20 Kannada and corresponding Tulu action verbs are randomly selected and these verbs along with their English translations are shown in Table 3.

2. 36 distinct suffixes belonging to present and future tenses in both Kannada and Tulu languages are selected to inflect the verbs and the samples of such suffixes are shown in Table 4.

3. Based on day-to-day conversation, 40 objects each for Kannada and Tulu languages which are compatible with the 20 action verbs mentioned in Table 4 are selected.

4. 4 different cases (*vibhakti*) are selected for object inflections for both Kannada and Tulu languages and these cases are shown in Table 4.

5. 10 sample words belonging to masculine and feminine genders and 4 pronouns (both including singular and plural) are randomly selected from the online resources.

6. Based on *vibhakti*/cases, PNG, tense, the intent of the verb and complexity of the verbal inflection, 14 categories are created for Kannada and 24 categories are created for Tulu as most of the words in Tulu change their spelling when a verb in its base form is inflected with the suffix/es.

7. A handcrafted implementation is carried out based on the rules set to get the unique combinations of subjects, objects, and verbs in Subject-Object-Verb (SOV) order. The sample rule sets along with sample subject, object, and verb combinations, for both Kannada and Tulu languages are shown in Table 5.

8. The order of subjects, objects, and verbs in unique SOV combinations is changed into OSV in order to get syntactical variation in the generated text.

With this arrangement, the proposed rule-based ATG with a set of 14 rules for Kannada and 24 rules for Tulu resulted, in 57,600 unique Kannada-Tulu parallel sentences. However, as both Kannada and Tulu are morphologically rich languages, it is very

Rule set	Subject	Object	Verb	Description
Kannada				

for all verbs:
if subject is 1st person singular
for suffix in list of suffixes:
get verbal_inflection
for all objects in list of vibhakti suffixes:
if object belongs to dative case and contain noun indicating place:
get object_inflection
return subject+object_inflection+verbal_inflection
	ನಾನು	ಪೇಟೆಗೆ	ಹೋಗುತ್ತೇನೆ	ನಾನು = '1st person singular'
				ಪೇಟೆಗೆ = 'noun dative case'
				ಹೋಗುತ್ತೇನೆ = 'present tense singular'

Tulu				

for all verbs:
if subject is 3rd person singular masculine
for suffix in list of suffixes:
get verbal_inflection
for all objects in list of vibhakti suffixes:
if object belongs locative case and contain noun indicating place:
get object_inflection
return subject+object_inflection+verbal_inflection
	ರಾಮೆ	ಇಲ್ಲ್ ಡ್	ಉಂತುವನ ದಾನ	ರಾಮೆ = '3rd person singular masculine'
				ಇಲ್ಲ್ ಡ್ = 'noun locative case'
				ಉಂತುವೆ = 'future tense singular'

Table 5: Sample subject, object, and verb combinations along with rule sets for Kannada and Tulu languages

difficult to define an exhaustive set of linguistic rules.

3.2.1 Evaluation of the text generated by ATG

Rule-based approaches can be useful for processing simple text and these approaches can be prone to errors, especially when dealing with morphologically rich and agglutinative languages due to rich verbal and object inflections (Vodolazova and Lloret, 2019). In addition, as there are no generalized rules for handling verbal and object inflections (Antony et al., 2012), syntactic and semantic evaluation of the text generated by the rule-based ATG system is very essential. The evaluation of the proposed ATG is carried out in two steps: Manual verification to find the wrong patterns in a selected set of sentences and automatic identification of these wrong patterns in the given set of sentences. The verification details are given below:

- Manual verification - of the generated text involves manually checking whether the generated sentences are syntactically and semantically correct or not. These sentences are then annotated with 'correct' or 'incorrect' labels depending on whether the sentence is correct or incorrect respectively. This annotation process involves 6 native speakers of the respective languages as annotators; 3 for each language. Guidelines provided to the annotators to carry out the annotations along with the sample Kannada and Tulu sentences followed by their English translations are shown in Table 6. Words that violate the rules in the incorrect sentences (both in Kannada and Tulu) are identified and listed, for creating wrong patterns and a sample of such wrong patterns are shown in Table 7. This process ensures the quality and accuracy of the generated content. Out of 57,600 Kannada-Tulu parallel sentences, 12,000 sentences containing all the verbal and object inflections are selected and these sentences are used for the manual verification so that all possible wrong patterns can be captured.

- Automatic evaluation - of the sentences is carried out by matching the wrong patterns (obtained during manual verification) automatically. Out of 57,600 sentences generated, 45,600 sentences are considered for automatic evaluation. If a sentence consists of wrong patterns, such a sentence is annotated as 'incorrect' otherwise it is annotated as 'correct'.

Out of 57,600 parallel sentences generated by ATG, 44,288 parallel sentences are found to be correct after evaluation. These sentence pairs/parallel sentences combined with the 72,974 manually translated sentence pairs/parallel sentences amounting to 1,17,262 sentence pairs/parallel sentences form the Kannada-Tulu parallel corpus and Table 8 shows the statistics of this corpus.

Label	Guideline	Sample Kannada sentence	Sample Tulu sentence	English translation	Description
Correct	Sentence with correct syntax and semantic information is tagged as 'correct'	ಅವರು ಮನೆಗೆ ಹೋಗುವುದಿಲ್ಲ	ಆರ್ ಇಲ್ಲ್ ಗ್ ಪೋಪುಜಿರ್	They are not going to home	The sentence conveys a complete information and is syntactically and semantically correct
Incorrect	Sentence with incorrect syntax and/or semantic information is tagged as 'incorrect'	ಮನೆದಲ್ಲಿ ಅವಳು ಹೋಗುತಾಳೆ	ಇಲ್ಲ್ ಟ್ ಆಲ್ ಪೋಪಲಾ	Does she go to home	In sample Kannada sentence, the word 'ಮನೆದಲ್ಲಿ' is syntactically wrong and in sample Tulu sentence, though the word 'ಇಲ್ಲ್ ಟ್' is syntactically correct, it used in a wrong context which is semantically incorrect

Table 6: Annotation guidelines along with sample Kannada and Tulu sentences followed by their English translations

Kannada	Tulu
['ಮಾರುಕಟ್ಟೆದಲ್ಲಿ','ಮಾರುಕಟ್ಟೆಕ್ಕೆ','ಶಾಲೆದಲ್ಲಿ','ಮನೆಕ್ಕೆ','ಪೇಟೆಕ್ಕೆ','ಪೇಟೆದಲ್ಲಿ','ಹಳ್ಳಿದಲ್ಲಿ','ಹಳ್ಳಿಕ್ಕೆ','ಶಾಲೆಕ್ಕೆ','ಮನೆದಲ್ಲಿ' 'ಕಾರ್ಯಕ್ರಮಗೆ', 'ಟಿವಿದಲ್ಲಿ', 'ಟಿವಿಕ್ಕೆ ', 'ಚಿತ್ರಕ್ಕೆ' , 'ತಿಂಡಿವನ್ನು']	['ಇಲ್ಲ್ ಕ್','ಇಲ್ಲ್ ಟ್',' ಹಳ್ಳಿಕ್','ಶಾಲೆಕ್',' ಹಳ್ಳಿಟ್', 'ಪೇಟೆಕ್', 'ಪೇಟೆಟ್', 'ಎದ್ದೆ ಗ್', 'ಟಿವಿಡ್ಡ್ ', 'ಟಿವಿಟ್ ', ಪರಂದ್ ಡ್', 'ಮೆಲ್ಲಟ್', 'ಜೋರಕ್ ', 'ಜೋರಡ್', 'ಗೊಬ್ಬುಗ್', 'ಬೇಲೆಟ್']

Table 7: Sample wrong patterns in Kannada and Tulu sentences

Languages	# of words	# of unique words	Average sentence length
Kannada	7,04,937	1,36,129	6
Tulu	7,81,603	1,33,568	7
Kan-Tul parallel sentences	**Train set**	**Test set**	
	1,14,762	2,500	
	Total	**1,17,262**	

Table 8: Statistics of Kannada-Tulu parallel corpus and details about Train and Test set

4 Neural Machine Translation Baselines

A set of encoder-decoder based NMT approaches including RNN with GRU (RNN+GRU) and LSTM (RNN+LSTM) units, BiRNN with GRU (BiRNN+GRU) and LSTM (Bi-RNN) units, transformers with GRU (transformers+GRU) and LSTM (transformers+LSTM) units, are experimented to set the benchmarks for the Kannada-Tulu parallel corpus. Further, sub-word tokenization is also explored in these approaches as it has exhibited promising results in similar research works (Hegde and Shashirekha, 2022). In order to implement NMT models, the following steps are used:

4.1 Pre-processing

Pre-processing plays a pivotal role in enhancing the quality of NMT output, as indicated by the research studies (Hegde et al., 2021b; Oudah et al., 2019). It not only eliminates the noise from the corpus, but also assesses the corpus semantically, encouraging the formation of sentence alignments.

4.2 Sub-word Tokenization

Sub-word tokenization is a popular tokenization technique where rare words are broken down into their most frequent words and represented by the sequences of bytes (Sennrich et al., 2015). The purpose of sub-word tokenization is to avoid OOV problems by analyzing a word as sub-words. In this work, BPE[3] - a popular technique for sub-word tokenization is employed.

4.3 Model Construction

NMT is a corpus-based approach for translation, leveraging Neural Networks (NN) to facilitate seamless processing of text from one language to another (Sutskever et al., 2014). It is widely regarded as the most suitable method for conducting sequence-to-sequence translation at the sentence level. The sequence-to-sequence architecture, also known as vanilla RNN, is fundamental for sequence prediction tasks. RNNs are beneficial for processing sequential data of variable lengths by utilizing a hidden state that retains information from previous time steps. This inherent capability allows RNNs to capture and leverage contextual information, enabling them to effectively model dependencies between words in tasks such as sentence-level translation. It comprises of two essential components: the encoder - takes the source language text as input and trans-

[3] https://bpemb.h-its.org/

Hyper-parameters	Values
word vector size	512
encoding layers	3
decoding layers	3
heads	8
learning rate	1.0
dropout	0.3
batch size	64
train steps	1,00,000
encoder type	transformer
rnn type	lstm
position encoding	True
optimization	sgd
check-point	10,000

Table 9: Hyper-parameters and their values used to configure Transformer+LSTM+BPE model

forms it into an intermediate representation, and the decoder - generates the output by utilizing the encoding vector and previously generated words (Neubig, 2017). This work utilizes different RNN architectures (LSTM and GRU) followed by an RNN variant called BiRNN (Schuster and Paliwal, 1997). Additionally, transformer - a self-attention-based NN is also employed with different RNN units (LSTM and GRU) (Vaswani et al., 2017).

5 Experiments and Results

NMT models for translating Kan-Tul and vice versa are implemented using the open-source NMT framework, OpenNMT-py[4]. This framework offers a structured encoder-decoder architecture enriched with attention mechanisms, allowing seamless handling of sequence-to-sequence prediction tasks. Several experiments are carried out involving the meticulous fine-tuning of hyperparameters to achieve optimal performances of the learning models. Variations in encoder types, such as the RNN, BiRNN, and transformer, are explored. Additionally, RNN architectures namely: LSTM and GRU are explored to identify the most suitable hyperparameters configurations. The optimal hyperparameters and their values for Transformer+LSTM+BPE models which exhibited better performance are shown in Table 9.

In assessing the performance of NMT models, BLEU scores and CHRF scores are chosen as the evaluation metrics. BLEU scores are calculated by counting matching word n-grams between the candidate translation and the reference text (Papineni et al., 2002), whereas CHRF measures the similarity between the candidate translation and one or more reference translations in terms of character-level n-grams (Popović, 2015). These approaches provide an automated means of evaluation similar to human assessment, enabling researchers to quantitatively measure the quality of MT models and make informed comparisons between different translation outputs.

The performances of the proposed models in terms of BLEU and CHRF scores are shown in Table 10 and the results illustrate that all the proposed NMT models exhibited considerably good BLEU and CHRF scores for both Kan-Tul and Tul-Kan translations. From Table 10, it is clear that models with GRU exhibited the lowest scores due to the simpler architecture of GRU which fails to capture the correct context of the words with complex structure. On the other hand, models with LSTM performed better, as LSTM architecture captures long-term dependencies. LSTM architectures are beneficial for dealing with complex data, which makes them suitable for handling morphologically rich agglutinative languages like Kannada and Tulu. BiRNN has shown slightly improved performance due to its ability to capture information from the past and future contexts, providing a more comprehensive understanding of the input sequence. Transformer models outperformed the other models because of their self-attention mechanism, which captures the context during translation. The combination of Transformer+LSTM+BPE model has shown a slight improvement over the other models because of the small Tulu vocabulary of 10,000 words in BPE.

The sample Kan-Tul translations obtained from RNN+GRU and Transformer+LSTM+BPE models along with the actual translations are shown in Table 11. It is clear that translations obtained from the RNN+GRU model have more unknown tokens compared to the translations obtained from the Transformer+LSTM+BPE model. Further, output 1 (RNN+GRU) has more <unk> (unknown) tokens compared to output 2 (Transformer+LSTM+BPE) as GRU is unable to capture the long-term dependencies because of its simple architecture with fewer parameters. In E.g. 1 and 3, output 2 has one unknown token. This may be due to the complexity of the word and rare occurrence of the word during training. Further, for output 2 in E.g. 2 and 4, all the source words are successfully translated, indicating efficiency of the Transformer+LSTM+BPE

[4] https://github.com/OpenNMT/OpenNMT-py

Models	Kannada-Tulu		Tulu-Kannada	
	BLEU	CHRF	BLEU	CHRF
RNN+GRU	0.098	0.196	0.10	0.20
RNN+LSTM	0.201	0.468	0.273	**0.543**
BiRNN+GRU	0.082	0.196	0.089	0.199
BiRNN+LSTM	**0.206**	**0.472**	**0.291**	**0.549**
Transformer+GRU	0.148	0.224	0.186	0.235
Transformer+LSTM	**0.238**	**0.483**	**0.301**	**0.562**
Transformer+GRU+BPE	0.205	0.257	0.216	0.311
Transformer+LSTM+BPE	**0.241**	**0.502**	**0.341**	**0.598**

Table 10: Performance of the baselines in terms of BLEU and CHRF scores

Kan-Tul translation					
E.g.	**Source sentence**	**Output 1 (RNN+GRU)**	**Output 2 (Transformer+ LSTM+BPE)**	**Actual translation**	**English translation**
1	ಇಂದು ಭಾರತ ಬಾಂಬ್ ಸ್ಫೋಟಿಸಿದೆ ಎಂದ	ಇನಿ <unk> <unk> <unk>	ಇನಿ ಭಾರತ ಬಾಂಬ್ <unk> ಪಂಡೆ	ಇನಿ ಭಾರತ ಬಾಂಬ್ ಸ್ಫೋಟಿಸಾಂಡ್ ಪಂಡೆ	He said that today India has exploded a bomb
2	ಅದೇ ರೀತಿ ಪೂರ್ವದಲ್ಲಿ ಬೆಳಕು ಕಾಣುತ್ತದೆ	ಅವ್ವೆ <unk> <unk> ಬೊಲ್ಪು ತೋಜುಂಡು	ಅವ್ವೆ ರೀತಿ ಪೂರ್ವಡ್ ಬೊಲ್ಪು ತೊಜಿಂಡು	ಅವ್ವೆ ರೀತಿ ಪೂರ್ವಡ್ ಬೊಲ್ಪು ತೊಜಿಂಡು	Similarly light appears in the east
Tul-Kan translation					
3	ಆಂಡ ಅವೆನ್ ಸರ್ಕಾರ ತಿರಸ್ಕಾರ ಮಲ್ತುಂಡು.	ಆದರೆ <unk> <unk>	ಆದರೆ ಅದನ್ನು ಸರ್ಕಾರ <unk>	ಆದರೆ ಅದನ್ನು ಸರಕಾರ ತಿರಸ್ಕರಿಸಿದೆ.	But, government has rejected that
4	ಒಂಜಿ ಪಾತ್ರೆಡ್ ಎಣ್ಣೆನ್ ಬೆಚ್ಚ ಮಲ್ಪರೆ ದೀಲೆ.	ಒಂದು <unk> <unk> <unk>	ಒಂದು ಪಾತ್ರೆಯಲ್ಲಿ ಎಣ್ಣೆಯನ್ನು ಬಿಸಿ ಮಾಡಲು ಇಡಿ.	ಒಂದು ಪಾತ್ರೆಯಲ್ಲಿ ಎಣ್ಣೆಯನ್ನು ಬಿಸಿ ಮಾಡಲು ಇಡಿ.	Heat oil in a pan.

Table 11: Kan-Tul and Tul-Kan translation samples obtained by the NMT models along with the actual translations followed by their English translations

model.

6 Conclusion and Future Work

This paper describes the construction of Kannada-Tulu parallel corpus for NMT and the NMT baselines to translate Kan-Tul and vice versa. Kannada-Tulu parallel corpus is constructed manually and also using ATG by incorporating linguistic rules. Using Kannada-Tulu parallel corpus, different NMT models with different encoder and decoder variants are implemented to translate Kan-Tul and vice versa. Among the proposed models, Transformer+LSTM+BPE model outperformed the other models with BLEU scores of 0.241 and CHRF scores of 0.502 for Kan-Tul translation. Further, the same model outperformed the other models for Tul-Kan translation with BLEU scores of 0.341 and CHRF scores of 0.598. Future research will examine hybrid NMT models that combine Kannada and Tulu language traits with appropriate pre-processing methods. Further, ATG will be enhanced by incorporating additional rules to address the current limitation of producing short-length sentences in future work.

References

PJ Antony, Hemant B Raj, BS Sahana, Dimple Sonal Alvares, and Aishwarya Raj. 2012. Morphological Analyzer and Generator for Tulu Language: A Novel Approach. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, pages 828–834.

PJ Antony, CK Savitha, and UJ Ujwal. 2016. Haar Eatures Based handwritten character recognition system for Tulu script. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 65–68. IEEE.

Shreya Banerjee. 2021. Determinants of Rural-urban Differential in Healthcare Utilization among the Elderly Population in India. In *BMC Public Health*, pages 1–18. BioMed Central.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A Survey on Data Augmentation for Text Classification. In *ACM Computing Surveys*, pages 1–39. ACM New York, NY.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Robert L Mercer, et al. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. MIT Press.

Wolfgang Butzkamm. 2003. We Only Learn Language Once. The Role of the Mother Tongue in FL Classrooms: Death of a Dogma. In *Language learning journal*, pages 29–39. Taylor & Francis.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the Conference of European Association for Machine Translation (EAMT)*, pages 261–268.

Asoka Chakravarthi and Bharathi Raja. 2020. Leveraging Orthographic Information to Improve Machine Translation of Under-resourced Languages.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Bernardo Stearns, Arun Kumar Jayapal, S Sridevy, Mihael Arcan, Manel Zarrouk, and John P McCrae. 2019. Multilingual Multimodal Machine Translation for Dravidian Languages utilizing Phonetic Transcription. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 56–63.

Sai Man Cheok, Lap Man Hoi, Su-Kit Tang, and Rita Tse. 2022. Crawling Parallel Data for Bilingual Corpus using Hybrid Crawling Architecture. In *Procedia Computer Science*, pages 122–127. Elsevier.

Asha Hegde, Shubhanker Banerjee, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Hosahalli Shashirekha, John Philip McCrae, et al. 2022a. 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.

Asha Hegde, Ibrahim Gashaw, and Shashirekha Hl. 2021a. Mucs@-machine Translation for Dravidian Languages using Stacked Long Short Term Memory. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 340–345.

Asha Hegde, Ibrahim Gashaw, and Shashirekha Hl. 2021b. Mucs@-Machine Translation for Dravidian Languages using Stacked Long Short Term Memory. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 340–345.

Asha Hegde and HL Shashirekha. 2020. MUCS@ Adap-MT 2020: low resource domain adaptation for Indic machine translation. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): Adap-MT 2020 Shared Task*, pages 24–28.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2022. KanSan: Kannada-Sanskrit Parallel Corpus Construction for Machine Translation. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 3–18. Springer International Publishing Cham.

Asha Hegde, Hosahalli Lakshmaiah Shashirekha, Anand Kumar Madasamy, and Bharathi Raja Chakravarthi. 2022b. A Study of Machine Translation Models for Kannada-Tulu. In *Congress on Intelligent Systems*, pages 145–161. Springer Nature Singapore Singapore.

Saméh Kchaou, Rahma Boujelbane, and Lamia Hadrich. 2023. Hybrid Pipeline for Building Arabic Tunisian Dialect-Standard Arabic Neural Machine Translation Model from Scratch. In *ACM Transactions on Asian and Low-Resource Language Information Processing*, pages 1–21. ACM New York, NY.

Amba Kulkarni and Madhusoodana Pai. 2019. Sanskrit Sentence Generator. In *Proceedings of the 6th International Sanskrit Computational Linguistics Symposium*, pages 1–13. Association for Computational Linguistics.

Lalita Lowphansirikul, Charin Polpanumas, Attapol T Rutherford, and Sarana Nutanong. 2022. A large English–Thai Parallel Corpus from the Web and Machine-Generated Text. In *Language Resources and Evaluation*, pages 477–499. Springer.

Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. 2022. A Review: Data Pre-processing and Data Augmentation Techniques. In *Global Transitions Proceedings*, pages 91–99. Elsevier.

Chenggang Mi, Lei Xie, and Yanning Zhang. 2022. Improving Data Augmentation for Low Resource Speech-to-Text Translation With Diverse Paraphrasing. In *Neural Networks*, pages 194–205. Elsevier.

Vandan Mujadia and Dipti Misra Sharma. 2022. The LTRC Hindi-Telugu Parallel Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3417–3424.

Farha Naznin, Shikhar Kumar Sarma, and Kishore Kashyap. 2023. Parallel Corpus Creation for NMT using Web Scraping and Filtering. In *2023 4th International Conference on Computing and Communication Systems (I3CS)*, pages 1–5. IEEE.

Graham Neubig. 2017. Neural Machine Translation and Sequence-to-Sequence Models: A Tutorial. In *arXiv preprint arXiv:1703.01619*.

Nafisa Nowshin, Zakia Sultana Ritu, and Sabir Ismail. 2018. A Crowd-Source Based Corpus on Bangla to English Translation. In *2018 21st International Conference of Computer and Information Technology (ICCIT)*, pages 1–5. IEEE.

Mai Oudah, Amjad Almahairi, and Nizar Habash. 2019. The Impact of Preprocessing on Arabic-English Statistical and Neural Machine Translation. In *arXiv preprint arXiv:1906.11751*.

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 of the Association for Computational Linguistics*, pages 311–318.

Mohammad Taher Pilevar, Heshaam Faili, and Abdol Hamid Pilevar. 2011. Tep: Tehran English-Persian Parallel Corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 68–79. Springer.

Maja Popović. 2015. chrF: Character N-gram F-score for Automatic MT Evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. In *Transactions of the Association for Computational Linguistics*, pages 145–162. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ....

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional Recurrent Neural Networks. In *IEEE transactions on Signal Processing*, pages 2673–2681. Ieee.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving Neural Machine Translation Models with Monolingual Data. In *arXiv preprint arXiv:1511.06709*.

Than Htut Soe, Frode Guribye, and Marija Slavkovik. 2021. Evaluating AI Assisted Subtitling. In *ACM International Conference on Interactive Media Experiences*, pages 96–107.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in neural information processing systems*.

Rita Tse, Silvia Mirri, Su-Kit Tang, Giovanni Pau, and Paola Salomoni. 2020. Building an Italian-Chinese Parallel Corpus for Machine Translation from the Web. In *Proceedings of the 6th EAI international conference on smart objects and technologies for social good*, pages 265–268.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in neural information processing systems*.

Tatiana Vodolazova and Elena Lloret. 2019. The Impact of Rule-Based Text Generation on the Quality of Abstractive Summaries. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1275–1284. INCOMA Ltd.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2021. Progress in Machine Translation. In *Engineering*. Elsevier.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized Data Augmentation for Low-Resource Translation. In *arXiv preprint arXiv:1906.03785*.

Shiwen Yu, Ting Wang, and Ji Wang. 2022. Data Augmentation by Program Transformation. In *Journal of Systems and Software*, page 111304. Elsevier.