Integrating Syntactic and Discourse Signals through Multi-Encoder Fusion in NMT for Low-Resource Indian Language Pairs

Sobha Lalitha Devi, Vijay Sundar Ram R, Pattabhi RK Rao

AU-KBC Research Centre
MIT Campus of Anna University,Chennai, India
sobha@au-kbc.org

Abstract

Neural Machine Translation (NMT) for low-resource Indian language pairs such as Hindi–Tamil and Tamil–Malayalam remains challenging due to morphological richness, syntactic divergence, and limited availability of high-quality parallel corpora. While Transformer-based architectures achieve strong performance in high-resource settings, they often struggle to model syntactic structure and discourse-level dependencies in low-resource scenarios, resulting in errors in agreement, word order, and pronoun translation. In this work, we propose a linguistically informed multi-encoder fusion framework that explicitly incorporates syntactic and discourse signals into NMT. Experiments conducted on Hindi–Tamil and Tamil–Malayalam parallel corpora demonstrate consistent improvements over strong Transformer baselines in BLEU and ChrF scores, along with gains in pronoun translation accuracy and agreement consistency. The results highlight the effectiveness of explicit linguistic integration for improving NMT in low-resource Indian language settings.

Keywords: Multi-Encoder Fusion, Linguistic Features, Neural Machine Translation, Low Resource Languages, Hindi, Tamil, Malayalam

1. Introduction

Neural Machine Translation (NMT) has significantly improved translation quality with the introduction of encoder–decoder architectures, especially with the Transformer model introduced in Attention Is All You Need (Vaswani, 2017). Despite these advances, NMT systems especially for Indian languages still struggle with:

- Syntactic ambiguities
- Long-distance dependencies
- Pronoun resolution and discourse consistency

Two major linguistic signals that can address these issues are:

- Part-of-Speech (POS) information – captures syntactic structure
- Anaphora resolution – resolves pronouns and coreference relations

Encoder fusion integrates these linguistic features directly into the NMT encoder to enhance contextual representation and improve translation quality.

Modern NMT systems typically use, Encoder and Decoder. An Encoder converts source sentence into contextual embedding. A Decoder generates target sentence based on encoded representation. The Transformer model in general uses:

- Multi-head self-attention
- Positional encoding
- Feed-forward layers

However, pure data-driven learning does not effectively capture explicit syntactic and discourse-level information.

POS information will help in disambiguation of homographs, improved word reordering and better syntactic alignment. Anaphora resolution identifies the antecedent for anaphor which it refers to. And in translation, especially for languages such as Hindi, Tamil and Malayalam, incorrect pronoun resolution leads to grammatical errors. Anaphora resolution helps in clarity and cohesion in discourse by resolving references to previously mentioned entities, which can effect gender and number agreement.

This paper is further organised as follows: Section 2 describes the related works in this area of research. Section 3 describes the data and its preparation. Section 4 describes the methodology. Section 5 describes experiments and results. Section 6 concludes the paper.

2. Related Works

Encoder fusion has emerged as an important architectural strategy in deep learning, particularly in sequence-to-sequence and multimodal models, where information from multiple encoder representations is combined to improve downstream performance. Gao et al (Gao, 2020) categorized fusion strategies into early, intermediate, and late fusion, depending on the stage at which representations are combined. Encoder fusion typically falls under intermediate fusion, where latent feature representations from one or more encoders are integrated to form a richer joint representation. Liu et al (2021) has presented work on improving translation output using encoder fusion technique for sequence-to-sequence model. They proposed a simple fusion method by fusing only the encoder embedding layers for the softmax layer. Their experiment revealed that this methodology learns more expressive bilingual word embedding by

Proceedings of WILDRE-8, pages 98–103
12 May 2026.

building between relevant source and target embeddings.

Das et al (2022) has used encoder fusion in their Personalized response selection system, where persona, emotion, and entailment information are fusion. They used the fusion strategies and concept-flow encoding to train a BERT-based model which outperforms the previous methods by 2.3% on original personas.

Recent survey work by Jiao et al (2024) emphasized that intermediate fusion at the encoder level allows models to preserve modality- or feature-specific information while still enabling effective interaction between representations.

Huang et al (2025) has used encoder fusion architectures to improve information retrieval tasks, where they have fused the text and image features.

Encoder fusion is utilized in many multi-model learning experiments, where separate encoders process different modalities such as text, images, or speech and fused together. Li and Tang (Li et al, 2024) provided a recent survey on multimodal alignment and fusion, highlighting attention-based and representation-level fusion methods that combine outputs from multiple encoders.

Building on these insights, this work adopts an encoder fusion framework to integrate linguistic features into neural machine translation. Unlike prior approaches that rely solely on final-layer representations or treat linguistic annotations as simple input embeddings, we explicitly fuse encoder representations to jointly model lexical and syntactic information. This approach is particularly relevant for morphologically rich and syntactically divergent language pairs, such as Hindi – Tamil and Tamil-Malayam, where explicit modeling of syntactic structure can help alleviate data sparsity and re-ordering challenges.

3. Data

There is a need of large number of parallel sentences to build a robust Neural Machine Translation (NMT) system. Thus data is very crucial in the development of NMT systems. Publicly available data are English centric. One of huge parallel corpus which is available is NLLB (Costa-Jussà, et. al., 2022). This is a large-scale multilingual bitext (parallel text) corpus created by Meta AI as part of their effort to support translation across many languages including low-resource ones. The dataset covers 148 English-centric language pairs (i.e., English ↔ X) and 1,465 non-English-centric pairs (i.e., X ↔ Y) using metadata mined bitext. (Costa-Jussà, et. al., 2022)

Samanantar is one of the largest publicly-available parallel corpora for Indic languages. It contains ~49.7 million sentence-pairs between English and 11 Indic languages (Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu). (Ramesh et. al., 2022)

Bharat Parallel Corpus Collection (BPCC) is another comprehensive and publicly available parallel corpus that includes both existing and new data for all 22 scheduled Indic languages. It is comprised of two parts: BPCC-Mined and BPCC-Human, totaling approximately 230 million bitext pairs. BPCC-Mined contains about 228 million pairs. (Jay et.al, 2023)

Most of the parallel sentences available in these corpora are between language X and English. Opus is an Open Parallel Corpora, which aggregates the publicly available parallel corpus and helps to build the parallel corpus for the available parallel corpora.

From the above corpora, we prepared parallel sentences for Hindi-Tamil and Tamil-Malayalam. We have collected 2987320 parallel sentences for Tamil to Hindi pairs. These parallel sentences are collected from Tamil-English and Hindi-English parallel sentences available in the above mentioned parallel data. Hindi to Tamil parallel sentences are collected by considering the Tamil and Hindi sentences which have common English sentence. Similarly we collected 1492138 Tamil-Malayalam parallel sentences. These parallel sentences were filtered using Comet and LaBSE scores. LaBSE (Language-agnostic BERT Sentence Embedding) is a multilingual BERT to produce language-agnostic sentence embeddings for 109 languages. This model combines masked language model (MLM) and translation language model (TLM) pretraining with a translation ranking task using bi-directional dual encoders. (https://github.com/bojone/labse). COMET (Crosslingual Optimized Metric for Evaluation of Translation) uses a pre-trained multilingual transformer encoder such as XLM-RoBERTa. It uses deep multilingual semantic representations. Each input sentence is encoded into contextual embeddings.

The encoder captures Cross-lingual semantic alignment, (https://unbabel.github.io/COMET/)

We selected the parallel sentences with scores greater than 0.85. We got 9,26,962 Tamil-Hindi parallel sentences and 3,73,165 Tamil-Malayalam parallel sentences.

274	நாட்	##களில்	இந்தவிலங்கு	##களின்	பால்	உற்பத்தி	280	கிலோ	##கி	##ரா	##ம்	ஆகும்	.
B-QT_QTC	B-N_NN	I-N_NN	B-DM_DMR	B-N_NN	I-N_NN	B-N_NN	B-N_NN	B-QT_QTC	B-N_NN	I-N_NN	I-N_NN	I-N_NN	B-V_VM_VF	B-RD_PUNC

Figure 1: Example 1 sentence after sub-word processing with each token aligned with the POS information

As mentioned earlier we intend to train the models with the linguistic information, namely, Part-of-Speech (POS) tags and Anaphor information. So we processed the source sentences with POS tagger and Anaphora resolution engine.

For any neural machine learning, tokenization process is very crucial step. In the following paragraphs we explain about the tokenization that is used in this work.

3.1 Tokenization

In Neural Machine translation, tokenization is beyond separation of words based on white space. Here sub-word tokenization is done using Statistical measures or linguistic features. This reduces the vocabulary size and increases the frequency of the tokens and improves the translation by handling rare words and unknown words. The sub-word tokenization is very beneficial to morphologically rich languages as vocabulary size is large compared to other languages due to productive inflectional and derivational suffixations. Sennrich et. al. (2016) presented the statistical word segmentation techniques which is based on simple character n-gram model and segmentation based on the byte pair encoding (BPE) comparison algorithm. BPE sub-word algorithm is one of the widely used sub-word tokenization algorithm.

The other sub-word tokenization algorithms include, WordPiece, SentencePiece, and language specific algorithms such as Mecab (a morphological analysis based Japanese tokenizer), Stanford Word Segmentation (a Chinese word segmenter based on Conditional Random Fields). Ram and Sobha (2023) has presented a comparative study on effectiveness of morphological based and BPE sub-word segmentation.

In this work, we have tokenized the data using Indic-tokenizer[1]. The major disadvantage with the BPE sub-word algorithm for Indian languages is that, it segments word into tokens which do not form valid letters of the language's alphabet set. Consider the following example 1.

Example 1:
Tamil Sentence:
"274 நாட்களில் இந்த விலங்குகளின் பால் உற்பத்தி 280 கிலோகிராம் ஆகும்"

After sub-word tokenization using BPE:

274 நாட@@ ்களில் இந்த விலங்க@@ ுகளின் பால் உற்பத்தி 280 க@@ ிலோ@@ கி@@ ராம் ஆகும் .

Here in the above example, the glyphs (mathras) such as ' ◌ு, ◌், ◌ி' should occur with the previous token to form a proper letter.

Example 2 shows the tokenized output for same sentence as in example 1, when it is tokenized using the Indic-tokenizer.

Example 2:

"274 நாட் ##களில் இந்த விலங்கு ##களின் பால் உற்பத்தி 280 கிலோ ##கி ##ரா ##ம் ஆகும் ."

Here the tokens have the valid letters.

Thus in the work presented here Indic-tokenizer is used in the NMT system development.

In the sub-word tokenization in both the methodologies the words are divided into sub-words. Now we need to align the POS and Anaphora information to the sub-words, which are originally for the wordforms before tokenization. We aligned the POS and Anaphora information with the subword tokenized sentences, by distributing the POS and Anaphora information assigned to word to its sub-words also. For this purpose the BIO format is used as in BIO format.
Consider the following example:

விலங்குகளின் (vilangkukaLin)
when processed with sub-word tokenization it is split into 'விலங்கு ##களின்'. The POS for this word is 'N_NN'. Here POS for the sub words are assigned as follows:

விலங்கு/B-N_NN ##களின்/I-N_NN

Similarly the same is followed for the Anaphora information also. The anaphora information is assigned to the sub-words as follows:

விலங்கு/B-Ante-3 ##களின்/I-Ante-3

Figure 1 shows the sentence in example 1 after sub-word processing in which each token is aligned with the POS information.

[1] https://github.com/sudarsun/indic-tokenizer

4. Methodology

Encoder fusion is a family of architectures where multiple encoders are combined to produce a single representation for downstream tasks. This shows up a lot in multilingual NLP, multimodal models, domain adaptation, and parameter-efficient fine-tuning.
The fusion mechanism operates as an additive multi-stream encoder where:

- Primary stream: Standard token embedding's capture semantic and lexical information
- Auxiliary streams: Separate embedding spaces for POS tags and chunk labels capture syntactic structure
- Fusion point: Early fusion at the embedding layer, before the self-attention stack

This differs from late fusion (combining features after encoding) or feature concatenation approaches by maintaining the original embedding dimensionality through projection.
In this work late fusion approach is used, having gating mechanism. In encoder fusion, multiple representations are combined:
Word embedding's, POS embedding's, Coreference/anaphora embedding's. These are fused within the encoder layers. The gating architecture mechanism works as follows:

$$H = \alpha\ H\{word\} + \beta\ H\{pos\} + \gamma\ H\{coref\}$$

Where weights α, β and γ are learned dynamically.

POS Embedding Layer:

Each POS tag is mapped to a dense vector:

$$E\_\{pos\} = Embedding(POS\_tag)$$

then, combined with word embedding:

$$E\_\{input\} = E\_\{word\} + E\_\{pos\}$$

Coreference Embedding Layer:

Coreference signals are encoded as:

- Binary features (is pronoun, is antecedent)
- Entity cluster embeddings
- Distance-based embeddings

These are fused into encoder representations.

The architecture used in this work adopts a multi-encoder design consisting of dedicated encoders for lexical tokens, Part-of-Speech (POS) sequences, and coreference annotations. Each encoder captures complementary aspects of linguistic structure—semantic content, syntactic function, and discourse-level referential relations—before their representations are integrated through attention-driven and learnable gating fusion mechanisms. By explicitly modeling token-level semantics, syntactic structure, and discourse-level coreference, this multi-encoder fusion framework produces richer contextual embedding's, facilitating discourse-coherent and syntactically accurate translations in low-resource Indian language pairs.

5. Experiments and Results

We evaluated the translations of the NMT models for both Hindi to Tamil and Tamil to Malayalam, using BLEU score (Papineni et al., 2002). We used Sacrebleu python library to calculate the BLEU scores. The results are presented in Table 1. The BLEU scores show that the model with POS and Anaphora/Coreference features integrated has improved by 3% the translation in both Hindi-Tamil and Tamil-Malayalam. We have developed two models viz.,

a) **Sys-1** – NMT trained with just the Parallel data using the indic tokenization (as explained in section 3.1)
b) **Sys-2 –** NMT trained with encoder fusion using indic tokenization.

SNo	**Details**	**Hindi to Tamil**		**Tamil to Malayalam**	
		BLEU	**chrF**	**BLEU**	**chrF**
1	Sys-1	26.23	52.33	31.46	61.87
2	Sys-2	29.88	57.97	35.66	66.42

Table 1: BLEU Score for Hindi - Tamil and Tamil - Malayalam

On analysis of the translation output from different experiments in both Hindi to Tamil and Tamil to Malayalam, our observations are as follows,

Sys-1: Translated sentences were complete but most of these translations were not the exact translation. Translations convey a different sense due to the choice of the verb generation.

There were also words omitted in the translation. Technical words and rare words were handled, but there were errors in it.

Sys-2: Clausal sentences were translated better than the systems. Verb phrase generation was exact, though there were errors.

Overall this output was observed to be more coherent and closer to human translation.

We have explained the translation output with examples in the further part of this section.

Ex 1.(HI to Tamil):
Hindi-Input:
म्यूटेशन आनुवंशकि में मलि सकते हैं.
(Mutations can be found in genetics.)
Tamil Translations:

Sys-1:பிறழ்வுகள் மரபணு மரபணுவில் இருக்கலாம்.
(Mutations can occur in the genetics genetics.)

Sys-2: பிறழ்வுகள் மரபணுவில் கிடைக்லாம்.
(Mutations can be found in genetics.)

In this example we observe that both Sys-1 and Sys-2 outputs are proper sentences but Sys-2 translation has better sense translation.

Ex2: Clausal Sentence
Hindi-Input:
इतहिास उस दौर से शुरू होता है जब लोग लखिने की कला जानते थे.
(History begins from the time when people knew the art of writing.)

Tamil Translations:
Sys-1: மக்கள் எழுதும் கலத்திலிருந்து தொடங்கும்போது வரலாறு தொடங்குகிறது.
(History begins from when people started writing.)

Sys-2: மக்கள் எழுதும் கலையை மக்கள் அறிந்த போது வரலாறு தொடங்குகிறது.
(History begins from the time when people knew the art of writing.)

In Ex 2, the Hindi sentence has a relative participle clause. The clause construction was correctly translated by the **Sys-2** system. It has generated the relative participle verb, 'அறிந்த' (aRintha).

Ex 3:
Hindi-Input:
ग्राउंड रखरखाव उपकरण, जैसे लॉन मोवर, रोलर्स, लाइम पाउडर मशीन, मार्कगि मशीन, घास काटने वाली तलवारें, दरांती, श्रब मास्टर, कटर .
(Ground Maintenance Equipments like Lawn Mower, Rollers, Lime Powder Machine, Marking Machine, Mower, Sickle, Shrub Master, Cutter etc.)
Tamil Translations:
Sys-1: தரை பராமரிப்பு சாதனங்கள், ஒவ்வாமைகள், லீக் தூள் இயந்திரங்கள், புல்வெளி இயந்திரங்கள், புல்வெளிகள், ஆப்பிரிக்கா, ஆப்பிரிக்கா, கேரட் மற்றும் பலர் போன்ற தரை பராமரிப்பு சாதனங்கள்.
(Ground care equipment like Ground care equipment, Allergies, Leek powder machines, Lawn machines, Lawns, Africa, Africa, Carrot and many others.)

Sys-2: புல் வெட்டும் இயந்திரம், ரோலர்கள், சுண்ணாம்பு பொடி இயந்திரங்கள், நின்மஞ்சள் அழைப்பான், குறியீட்டு இயந்திரம், புல் வெட்டும் வாள்கள், நறுமணங்கள், சிரோமாஸ்டர், கடர் போன்ற தரை பராமரிப்பு உபகரணங்கள் அகும்.
(Ground care equipment like such as Lawn care equipment such as lawnmowers, lawnmowers, trucks, lime powders, machines, machines, machines, machines, mowers etc.)

In example 3, the Hindi sentence has series of noun phrases. **Sys-1** has generated output with many words which are not in the input sentences such as 'Africa', 'Carrot' etc. **Sys-2**, most of the noun phrases is translated and also the structure of the Tamil sentence is generated properly. This shows that Sys-2 has better performance than Sys-1. It requires little more data in the training.

We have also performed human evaluation using three human evaluators. The evaluation metrics used was Fluency **(F)** and Comprehensibility **(C).** Human evaluators had scored in a scale of 1-10 (1 indicates the lowest and 10 the highest score). In the table 2 we present the average human evaluation scores for both **Sys1** and **Sys2.**

SNo	Details	Hindi to Tamil		Tamil to Malayalam	
		F	C	F	C
1	Sys -1	62.66	69.88	67.65	71.34
2	Sys-2	71.27	79.45	78.85	81.35

Table 2: Human Evaluation Scores for the Sys1 and Sys2

6. Conclusion

We have presented our work in building Neural Machine Translation system in which we incorporated the syntactic and semantic features into the model development through encoder fusion gated mechanism. We have compared our Encoder Fusion NMT with NMT without Encoder Fusion. Hindi is an Indo-Aryan language. Malayalam and Tamil are Dravidian languages. All the three languages are morphologically rich language. And Tamil and Malayalam are highly agglutinative. The languages have different semantic and syntactic features such as the pronominals usage, PNG agreements etc. In our experiments we have observed that the encoder fusion has significant improvement. We have obtained 3% improvement. In future we plan to conduct ablation studies and also compare with LLMs.

7. Acknowledgments

This work is part of the research project titled "Discourse Integrated Dravidian Language to Dravidian Language Macine Translation (DL-DiscoMT)" funded under National Language Translation Mission (NLTM), Bhashini by Ministry of Electronics and Information Technology (MeitY), Government of India.

8. Bibliographical References

Castor, A. and Pollux, L. E. (1992). The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

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.

Costa-Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Mejía González, G., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., amp; Wang, J. (2022). "No Language Left Behind (NLLB): Scaling Human-Centered Machine Translation". arXiv preprint arXiv:2207.04672 https://doi.org/10.48550/arXiv.2207.04672

Souvik Das, Sougata Saha, and Rohini K. Srihari. (2022). "Using Multi-Encoder Fusion Strategies to Improve Personalized Response Selection". In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 532–541, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Gala, Jay, Pranjal A. Chitale, A. K. Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M., Janki Atul 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." Transactions on Machine Learning Research.

Wang, Yang. (2020). Deep multi-modal data analytics: Collaboration, rivalry and fusion. arXiv preprint arXiv:2006.08159.

Li, Songtao and Hao Tang. (2024). Multimodal alignment and fusion: A survey. arXiv preprint arXiv:2411.17040.

Liu, Xuebo and Wang, Longyue and Wong, Derek F. and Ding, Liang and Chao, Lidia S. and Tu, Zhaopeng. (2020). "Understanding and Improving Encoder Layer Fusion in Sequence-to-Sequence Learning.", Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP) https://arxiv.org/abs/2012.14768

Vijay Sundar Ram and Sobha Lalitha Devi. (2023). "Hindi to Dravidian Language Neural Machine Translation System". In: Proceedings of Recent Trends in Natural Language Processing (RANLP), 2023.

Gao, Jing, Peng Li, Zhikui Chen, and Jianing Zhang. (2020). A survey on deep learning for multimodal data fusion. Neural Computation 32(5):829–864.

Jiao, Tianzhe, Chaopeng Guo, Xiaoyue Feng, Yuming Chen, and Jie Song. (2024). A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. Computers, Materials & Continua 80(1):1–35.