Data Augmentation for English–Hindi Parallel Corpora using Parse Trees and Large Language Models

Mrunal Kumbhare (24MCS004)

Under the Guidance of: **Dr. Arun Kumar Yadav**

Department of Computer Science & Engineering

National Institute of Technology, Hamirpur (H.P.)

Table of Contents

Introduction

- India is one of the most linguistically diverse nations, with **121 languages and over 19,500 dialects**, according to the *2011 Census of India* [1].
- Because of this diversity, communication between people speaking different languages is often difficult.
- In today's digital world, we need translation systems that can help people understand and share information in different languages.
- Machine Translation (MT) is one such technology that helps translate text from one language to another.
- But to train these MT systems, we need a large amount of good quality data, known as parallel corpora (same text in two languages).

Introduction (contd.)

- For the English–Hindi language pair, limited data availability makes translation quality poor, making it a low-resource pair compared to English–French or English–German.
- Data augmentation can help by creating more training data from the existing one.
- This work aims to improve the quality and amount of English–Hindi data for better translation results.

Motivation

- India has hundreds of languages, but most online content and digital tools are available only in English or a few major Indian languages.
- This creates a language barrier, especially for people in rural and regional areas who prefer using their local language.
- This work was motivated by the idea of making technology accessible to everyone, no matter what language they speak.
- Many government messages, educational materials, and health information are not easily available in regional languages, which limits people's awareness and opportunities.

Motivation (contd.)

- People from non-English backgrounds often struggle to use online services such as digital banking, healthcare portals, and e-learning platforms, leading to digital inequality.
- By improving machine translation for English–Hindi, this work can help make information more accessible and understandable for a larger audience.
- The goal is not just technical improvement, but also to bridge the communication gap and support digital accessibility across language boundaries.

Literature Review

Reference & Year	Language	Key Findings	Limitations
[2] (2025)	English, Chinese, German	LLM back generation + contrastive span learning improve cross-domain parsing; achieves SOTA results.	Evaluated only on English; high GPT-4 generation cost; limited exploration of multilingual and end-to-end LLM integration.
[3] (2025)	English–Hindi	Used character-level Transformer with 5 phrase-based data augmentation; improved Hindi-English translation for low-resource settings.	Limited to fixed 5 phrase types; lacks deeper syntactic context; tested mainly on WMT14 and Samanantar datasets.
[4] (2024)	English, German, French, Chinese, Russian, etc. (12 languages)	Proposed XLM-DM — an unsupervised MNMT framework using XLM-R and bilingual dictionaries with word & sentence-level alignment. Outperformed baselines, showing strong zero-shot, few-shot, and domain-transfer abilities.	Needs quality bilingual dictionaries and monolingual data; limited testing on very low-resource or distant languages; computationally heavy for large-scale use.

Literature Review (contd.)

Reference & Year	Language	Key Findings	Limitations
[5] (2024)	11 Indic languages (incl. English–Hindi)	Created the largest multilingual parallel corpus for Indian languages; enables better MT model training.	Data still uneven across languages; limited domain diversity.
[6] (2024)	11 Indic languages – Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu	Built MNMT models for 11 Indic languages; English pivot and transliteration improved BLEU; related-language grouping helped West Indo-Aryan pairs.	Grouping had negative or inconclusive impact for East Indo-Aryan and Dravidian groups; performance limited by small datasets and low-resource languages like Assamese.
[7] (2023)	English to 15 Indic (Hindi, Bengali, Tamil, Telugu, etc.)	Transformer-based MNMT (15 Indic-Eng pairs) using shared encoder-decoder, backtranslation & language similarity to boost low-resource translation.	Residual noise in corpora; limited improvement from domain adaptation; zero-shot & unsupervised methods not explored.

Literature Review (contd.)

Reference & Year	Language	Key Findings	Limitations
[8] (2022)	English–Manipuri	Combined self-training, back-translation, and noise to boost English–Manipuri NMT; achieved +4.5 BLEU gain.	Training pipeline not fully optimized; limited monolingual data use; tested only on Bengali script.
[9] (2021)	7 Indic languages — Hindi, Bengali, Tamil, Telugu, Malayalam, Marathi, Gujarati	Proposes a multilingual NMT system using shared encoders and transfer learning to enhance translation quality.	Uneven accuracy across languages; limited testing on very low-resource pairs.

Summary of the Literature

- Recent research in machine translation and parsing focuses on combining grammar knowledge with data-driven learning.
- Techniques such as LLM-based parse tree generation and contrastive learning have been used to improve understanding of sentence structure across multiple languages.
- Other approaches like character-level translation models and phrase-based data augmentation have helped improve translation quality for low-resource pairs like English–Hindi.
- Together, these studies show that combining syntactic structure with data augmentation and additional phrase-level data can make translation models more accurate and context-aware.
- However, existing works still do not combine both grammar-based and data-driven methods in one system, and they rarely handle idioms or meaning-based translation in low-resource bilingual languages like Hindi–English.

Research Gap

- Most existing Machine Translation (MT) systems focus on high-resource languages like English–French or English–German [4].
- English–Hindi and other Indic pairs still have very few good-quality parallel datasets [5].
- Traditional data augmentation methods (like synonym replacement or back-translation) often fail to preserve syntax and meaning.
- Linguistic features such as parse trees or grammatical structures [3] are rarely used in existing augmentation research.
- Few studies explore the use of Large Language Models (LLMs) [2] for syntactically controlled data augmentation in low-resource MT.
- There is a need for a systematic approach that combines structure (syntax - parse trees) and meaning (semantics - LLMs) to improve the quality and diversity of bilingual data.

Problem Statement

- Machine Translation between English and Hindi faces challenges due to a shortage of high-quality parallel data.
- Existing English–Hindi translation datasets are limited, and current data augmentation methods fail to preserve meaning and structure, reducing translation quality.
- There is a need for a syntactically guided approach that can generate grammatically correct and meaningful data to enhance translation performance.

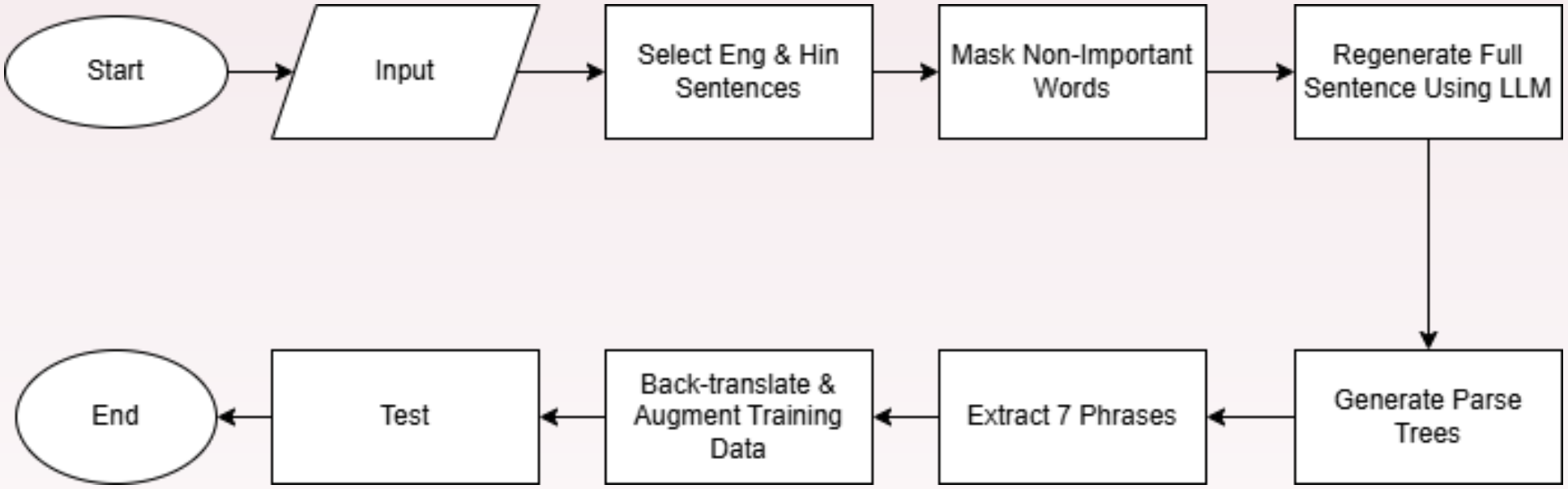
Research Objective

- To build a syntax-based data augmentation method for English–Hindi translation using parse trees to maintain sentence structure.
- To apply Large Language Models (LLMs) for generating fluent and meaningful new sentences.
- To use phrasal sentences to improve translation quality and handle idiomatic expressions more accurately.

Proposed Work

- This work aims to improve English–Hindi translation data through smart data augmentation.
- The idea is to use parse trees to understand the sentence structure and guide the generation of new sentences.
- Important keywords are kept unchanged, while the rest of the sentence is rephrased using a Large Language Model (LLM).
- The LLM generates new, grammatically correct and meaningful English sentences that have the same meaning as the original ones.
- These new sentences are then paired with the same Hindi translations to expand the dataset.

Proposed Work (contd.)



Work In Progress

- **Literature Review:** Conducted detailed study of NMT and constituency parsing techniques, focusing on phrase-based augmentation and LLM-driven syntactic learning.
- **Base Paper Analysis:** Reviewed and compared two base papers — *Character-Level Encoding based NMT* and *Cross-Domain Constituency Parsing via LLMs*.
- **Problem Identification:** Identified lack of an integrated syntactic-semantic framework for Hindi-English translation, especially in idiomatic and low-resource contexts.
- **Pipeline Design:** Designed combined methodology integrating 7 phrase-type extraction, GPT-based meaningful phrase generation, and mBART-50 back-translation.

Work Done So Far (Part 2)

- **Dataset Preparation:** Collected and preprocessed English–Hindi parallel corpora (WMT14, Samanantar) for sentence- and phrase-level augmentation. Also compiled a dedicated **Idioms & Phrases dataset** to address ambiguity in idiomatic expressions.

- **Implementation Progress:** Completed English-side pipeline up to the **data augmentation stage**, including:
  - Masked non-key words and back-generated English sentences using Groq's GPT-OSS-20B (a GPT-based model).
  - Generated constituency parse trees for English sentences using the same Groq GPT-OSS-20B model.
  - Extracted phrases using both fixed 7 phrase types and dynamic meaningful phrases.
  - Translated extracted English phrases to Hindi using IndicTrans2.
  - Augmented NMT training data with generated English–Hindi phrase pairs.

- **Next Steps:** Begin Hindi-side pipeline development, integrate idiom–phrase dataset, and train the English–Hindi Transformer-based NMT model using subword-level encoding.

Results

	en_clean	en_masked	en_backgen	en_parse_backgen
0	i went to the spot, saw a mesh of overhanging wires.	<extra_id_0> went <extra_id_0> <extra_id_0> spot, saw <extra_id_0> mesh <extra_id_0> overhanging wires.	I went to the spot, saw a mesh covering overhanging wires.	(S (NP (PRP I)) (VP (VBD went) (PP (TO to) (NP (DT the) (NN spot))) (, ,) (VP (VBD saw) (NP (DT a) (NN mesh)) (VP (VBG covering) (NP (JJ overhanging) (NNS wires))))) (. .))
1	but they are just the beginning.	<extra_id_0> <extra_id_0> <extra_id_0> just <extra_id_0> beginning.	We we we are just we beginning.	(S (NP (PRP We)) (VP (VBP are) (ADVP (RB just)) (VP (VBG beginning) (NP (PRP We)))) (. .))
2	rajiv collapsed on the spot.	rajiv collapsed <extra_id_0> <extra_id_0> spot.	Rajiv collapsed on the same spot.	(S (NP (NNP Rajiv)) (VP (VBD collapsed) (PP (IN on) (NP (DT the) (JJ same) (NN spot)))) (. .))
3	this usually has to be purchased.	<extra_id_0> usually <extra_id_0> <extra_id_0> <extra_id_0> purchased.	Items usually are carefully are thoroughly are rarely purchased.	(S (NP (NNS Items)) (VP (ADVP (RB usually)) (VP (VBP are) (ADVP (RB carefully)) (VP (VBP are) (ADVP (RB thoroughly)) (VP (VBP are) (ADVP (RB rarely)) (VP (VBN purchased))))) (. .))
4	name of post: junior assistant	<extra_id_0> <extra_id_0> post: junior assistant	We are currently looking to hire a junior assistant.	(S (NP (PRP We)) (VP (VBP are) (ADVP (RB currently)) (VP (VBG looking) (PP (TO to) (VP (VB hire) (NP (DT a) (JJ junior) (NN assistant))))) (. .))

Results (contd.)

✅ Loaded cleaned phrases: 37

	sentence_id	phrase_type	phrase_text	phrase_length	method
0	0	PP	to the spot	3	fixed
1	0	VP	covering overhanging wires	3	fixed
2	0	NP	a mesh	2	fixed
3	1	VP	are just beginning We	4	fixed
4	2	PP	on the same spot	4	fixed

- **Phrases we've choosen:** Noun, Verb, Preposition, Adjective, Adverb, Quantifier phrase, Interjection

Results (contd.)

✅ Translation complete!

	phrase_text	hi_phrase
0	to the spot	स्थान पर
1	covering overhanging wires	ओवरहंगिंग तारों को कवर करना
2	a mesh	एक जाल
3	are just beginning We	हम अभी शुरू कर रहे हैं
4	on the same spot	एक ही स्थान पर
5	Rajiv	राजीव
6	rainfall Light to moderate	वर्षा हल्के से मध्यम
7	expected in some places	कुछ स्थानों पर अपेक्षित
8	the Hyderabad Cricket Association	हैदराबाद क्रिकेट संघ
9	his nomination	उनकी नियुक्ति

Results (contd.)

✅ Average semantic similarity across 25 samples: 0.905

	phrase_text	hi_phrase	semantic_similarity
26	here	यहाँ	0.979532
11	recently	हाल ही में	0.973157
24	marked on the map	नक्शे पर चिह्नित	0.961509
16	the official announcement	आधिकारिक घोषणा	0.950986
21	including Afghanistan	अफगानिस्तान सहित	0.947119
27	said that he he	कहा कि वह वह	0.943445
19	of middle eastern countries	मध्य पूर्व के देशों के	0.942039
4	on the same spot	एक ही स्थान पर	0.937680
36	Unknown	अज्ञात	0.935024
6	rainfall Light to moderate	वर्षा हल्के से मध्यम	0.915713

References

[1] O. of the Registrar General and I. Census Commissioner, "C-16: Population by mother tongue, census of india 2011," 2018.

[2] P. Guo, M. Zhang, J. Li, M. Zhang, and Y. Zhang, "Contrastive learning on llm back generation treebank for cross-domain constituency parsing," *arXiv preprint arXiv:2505.20976*, 2025.

[3] D. Rathod, A. K. Yadav, M. Kumar, and D. Yadav, "Character-level encoding based neural machine translation for hindi language," *Neural Processing Letters*, vol. 57, no. 2, p. 23, 2025.

[4] Y. Shen, W. Bao, G. Gao, M. Zhou, and X. Zhao, "Unsupervised multilingual machine translation with pretrained cross-lingual encoders," *Knowledge-Based Systems*, vol. 284, p. 111304, 2024.

[5] G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. Jobanputra, R. Ak, A. Sharma, S. Sahoo, H. Diddee, D. Kakwani, N. Kumar, *et al.*, "Samanantar: The largest publicly available parallel corpora collection for 11 indic languages," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 145–162, 2022.

[6] S. Bala Das, D. Panda, T. Kumar Mishra, B. Kr. Patra, and A. Ekbal, "Multilingual neural machine translation for indic to indic languages," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 5, pp. 1–32, 2024.

References

[7] S. Bala Das, A. Biradar, T. Kumar Mishra, and B. Kr. Patra, "Improving multilingual neural machine translation system for indic languages," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 6, pp. 1–24, 2023.

[8] S. M. Singh and T. D. Singh, "Low resource machine translation of english–manipuri: A semi-supervised approach," *Expert systems with applications*, vol. 209, p. 118187, 2022.

[9] K. K. Gupta, S. Sen, R. Haque, A. Ekbal, P. Bhattacharyya, and A. Way, "Augmenting training data with syntactic phrasal-segments in low-resource neural machine translation," *Machine Translation*, vol. 35, no. 4, pp. 661–685, 2021.

Thank You!

Questions, comments, or feedback are welcome.

Mrunal Kumbhare

24MCS004

Department of Computer Science, NITH

24mcs004@nith.ac.in