Graph-Assisted Culturally Adaptable Idiomatic Translation for Indic Languages

Pratik Rakesh Singh, Kritarth Prasad, Mohammadi Zaki and Pankaj Wasnik
Media Analysis Group, Sony Research India
{pratik.singh, kritarth.prasad, mohammadi.zaki, pankaj.wasnik}@sony.com

Abstract

Translating multi-word expressions (MWEs) and idioms requires a deep understanding of the cultural nuances of both the source and target languages. This challenge is further amplified by the one-to-many nature of idiomatic translations, where a single source idiom can have multiple target-language equivalents depending on cultural references and contextual variations. Traditional static knowledge graphs (KGs) and prompt-based approaches struggle to capture these complex relationships, often leading to suboptimal translations. To address this, we propose IdiomCE, an adaptive graph neural network (GNN) based methodology that learns intricate mappings between idiomatic expressions, effectively generalizing to both seen and unseen nodes during training. Our proposed method enhances translation quality even in resource-constrained settings, facilitating improved idiomatic translation in smaller models. We evaluate our approach on multiple idiomatic translation datasets using reference-less metrics, demonstrating significant improvements in translating idioms from English to various Indian languages.

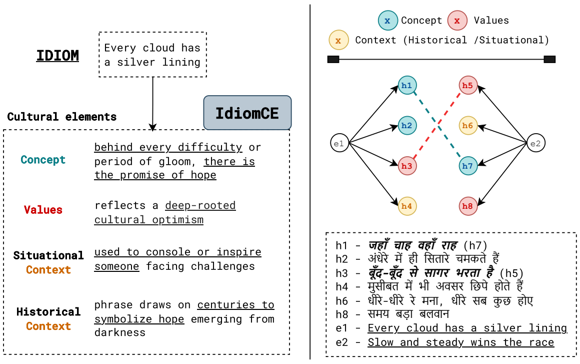
IDIOM
Every cloud has a silver lining
IdiomCE
Cultural elements
Concept
behind every difficulty or period of gloom, there is the promise of hope
Values
reflects a deep-rooted cultural optimism
Situational Context
used to console or inspire someone facing challenges
Historical Context
phrase draws on centuries to symbolize hope emerging from darkness
x Concept
x Values
x Context (Historical /Situational)
h1
h2
h3
h4
h5
h6
h7
h8
e1
e2
h1 - जहाँ चाह वहाँ राह (h7)
h2 - अंधेरे में ही सितारे चमकते हैं
h3 - बूँद-बूँद से सागर भरता है (h5)
h4 - मुसीबत में भी अवसर छिपे होते हैं
h6 - धीरे-धीरे रे मना, धीरे सब कुछ होए
h8 - समय बड़ा बलवान
e1 - Every cloud has a silver lining
e2 - Slow and steady wins the race

Figure 1: An example of cultural enhanced graph with different cultural elements: Concepts, Values, Context (Historical/Situational) and how we can create relationship among source and target nodes using their cultural elements.

1 Introduction

In linguistic terms, *idiom* is a *multi-word expression* (MWE) whose meaning cannot be derived from the literal meanings of its individual parts. Idioms have key properties such as noncompositionality, fixedness, and cultural specificity (Nunberg et al., 1994). They are integral to everyday language, enhancing expressiveness and communicative vividness. They often originate from diverse cultural, historical, and situational contexts, making them unique to specific languages or regions (Vula and Tyfekçi, 2024; Yagiz and Izadpanah, 2013).

With advancements in large language models (LLMs), neural machine translation (NMT) has significantly improved in handling complex linguistic phenomena, which led to research interest in complex linguistic tasks such as translating idioms across multiple languages (Li et al., 2023a; Rezaeimanesh et al., 2024a; Castaldo and Monti, 2024). However, despite these advancements, idiomatic translation remains a major challenge due to the inherent properties of idioms. Traditional NMT systems, both statistical and neural, struggle with noncompositionality, as they primarily process text at the word or phrase level rather than capturing an idiom's holistic meaning. This often leads to literal translations, distorting the intended meaning of the source text (Baziotis et al., 2023; Raunak et al., 2023; Dankers et al., 2022).

Recent efforts to address idiomatic translation have primarily relied on (1) idiom dictionary-based substitution (Salton et al., 2014) and (2) prompting techniques, such as chain-of-thought (CoT) reasoning or explicitly providing figurative meanings in prompts (Castaldo and Monti, 2024; Li et al., 2023b; Rezaeimanesh et al., 2024b). Although these methods have shown improvements in idiomatic translation, they still fail to overcome key challenges. As shown in Figure 1, these methods often overlook cultural factors that shape idioms and

Findings of the Association for Computational Linguistics: ACL 2025, pages 7029–7044
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

influence their mappings across languages (*Challenge I*). Additionally, they fail to address the one-to-many nature of idioms, where a single source-language idiom may have multiple valid translations in the target language, with the optimal choice depending on the source sentence's context (Rezaeimanesh et al., 2024a) (*Challenge II*). Moreover, knowledge graph (KG)-based approaches are inherently constrained by the availability of idiom resources, leading to translation gaps when encountering idioms not present in the KG (Peng et al., 2023) (*Challenge III*). These challenges pose a critical research question:
How can cultural nuances be effectively integrated into many-to-many idiomatic translation to enhance model performance?

To address this challenge, one possible approach is to first analyze the cultural dependencies of idioms and identify the specific cultural elements that shape idiomatic expressions across languages. Recent studies in NLP (Liu et al., 2024) (Pawar et al., 2024) introduce a comprehensive taxonomy of cultural and sociocultural elements, highlighting the need for culturally adaptive models as well as efforts to incorporate cultural awareness. However, even with a structured understanding of these cultural elements, capturing their intricate relationships and effectively leveraging them for one-to-many idiomatic translation remains a significant challenge.

This paper introduces **IdiomCE**, an inductive graph-based approach that models the relationships between source and target idioms by leveraging complex cultural element mappings, as illustrated in Figure 1, where source is an English idiom and target are Hindi idioms. Using link prediction, our method facilitates one-to-many idiomatic translation while preserving cultural relevance across languages. Furthermore, IdiomCE is adaptable, enabling the translation of unseen idioms by leveraging the inductive capabilities of GNNs, effectively addressing the limitations of noisy and limited knowledge bases. Our key contributions are summarized as follows:

- We propose a *cultural element-based data creation* method that generates multiple target idioms for a given source idiom.
- We develop an Inductive GNN trained on this graphical data, leveraging link prediction to enable one-to-many idiomatic translation (addressing *Challenge I* and *II*).
- We design an adaptable pipeline that extends to unseen idioms using the inductive capabilities of GNNs (addressing *Challenge III*).
- Using English as a pivot language, we extend our approach to facilitate idiomatic translation across Indic languages without needing to train GNN models between all possible pairs of languages.

2 Related work and Motivation

2.1 Related Works

Idiomatic Text Translation. Previous studies have explored various strategies to enhance NMT performance for idiomatic translation. (Salton et al., 2014) introduced a substitution-based method, where source-side idioms are replaced with their literal meanings before translation and reinstated post-translation to improve accuracy. (Zaninello and Birch, 2020) demonstrated that augmenting training data with idiomatic translations enhances model performance on both source and target sides. Beyond direct translation techniques, researchers have focused on learning non-compositional embeddings and automatically identifying idioms, as explored by (Weller et al., 2014), (Hashimoto and Tsuruoka, 2016), and (Tedeschi et al., 2022). More recently, prompting techniques and Chain-of-Thought (CoT) reasoning have been investigated in Large Language Models (LLMs) for idiomatic translation (Castaldo and Monti, 2024; Rezaeimanesh et al., 2024b). IdiomKB (Li et al., 2023a) further introduced a contextual approach, using figurative meanings as additional context to improve translation quality in LLMs.
Idiomatic Translation for Indic languages. Indic languages exhibit significant linguistic diversity and deeply rooted cultural nuances, making idiomatic translation a complex challenge. Despite this, research on idiomatic translation in the Indic language setting remains limited. (Shaikh, 2020) proposes Idiom Identification using grammatical rule based approach.(Modh and Saini, 2020) proposes a identification of Gujarati idioms and translation of them using contextual information. (Agrawal et al., 2018) present a multilingual parallel idiom dataset encompassing seven Indian languages and English. While these studies offer valuable contributions, the challenge of many-to-many idiomatic translation across Indic languages remains largely under-explored.

2.2 Motivation

Motivation for Cultural significance in Idioms. As discussed previously, most of the past studies either use a dictionary-based approach for idiom translation, which is a one-to-one mapping, or provide *figurative meaning* of the idiomatic expression for meaningful translation. Although these approaches appear to perform well, they fail to account for the cultural dependency of idioms, which is deeply embedded within them. This raises the question of how idioms can be effectively mapped from one language to another while considering this cultural dependency. Cultural dependency can be linked to various features, as discussed in (Liu et al., 2024) and (Pawar et al., 2024). Identifying these features that influence translation between languages can contribute to the development of more culturally appropriate idiomatic mappings from a source language to a target language.

Motivation for Using GNNs. Using a static Knowledge Graph (KG) or dictionary-based approach poses several challenges, which a Graph Neural Network (GNN)-based architecture can effectively address:

Limited Generalization. KGs store only predefined idiomatic translations as edges between nodes, making them incapable of inferring translations for new idioms unless explicitly added. In contrast, GNNs learn graph patterns, enabling them to predict idiomatic translations even for unseen idioms.

Lack of Semantic Connectivity. KGs treat nodes independently, failing to capture relationships between idioms with similar meanings unless explicitly modeled. GNNs leverage neighborhood structures and embeddings, allowing them to infer new translations by recognizing semantic similarities.

Polysemy Handling. KGs require multiple nodes to represent idioms with multiple meanings, increasing complexity. GNNs disambiguate meanings using context, leveraging neighborhood information and learned representations to differentiate between senses based on connectivity.

3 Methodology

In this section, we first present the problem statement followed by the training and inference of our methodology, which we call **IdiomCE**.

3.1 Problem Formulation:

We address the challenge of replacing idioms in a source language with culturally aware and contextually appropriate multi-word expressions in the target language. Let $\mathcal{S}$ and $\mathcal{T}$ denote the sets of graph nodes representing source and target idioms, respectively. The combined set $\mathcal{S} \cup \mathcal{T}$ defines the node set $\mathcal{V}$ in our framework, where each node $v \in \mathcal{V}$ corresponds to an idiom.

Each Idiom v is embedded with cultural elements, reflecting its historical, situational, or value-based significance, indicating its relevance to a specific language. Our goal is to identify the most relevant set of target-language idioms $\{\bar{v} : \bar{v} \in \mathcal{T}\}$ that correspond to a given source-language idiom v. We denote this relationship with an edge $e_{v,\bar{v}}$. Let the set of all such edges be $\mathcal{E} \equiv \{e_{v,\bar{v}} : v \in \mathcal{S}, \bar{v} \in \mathcal{T}\}$ Once we construct or estimate the graph $\mathcal{G} \equiv (\mathcal{V}, \mathcal{E})$, we use it to generate translations that are both contextually and culturally relevant. Given a sentence in the source language, our approach leverages this graph $\mathcal{G}$ to produce a culturally and semantically appropriate idiomatic translation in the target language.

3.2 Training

In this section, we outline the process of constructing the initial dataset for training our IdiomCE encoder and decoder, followed by the training methodology. An overview of the entire training process is illustrated in Figure 2.

GNN Dataset Formation. We begin by extracting idioms from the collected dataset, as detailed in Section 4 (Datasets), and obtain monolingual idiom sets for each language. For each idiom, we extract three key cultural elements: *Concepts*, *Values*, and *Situational and Historical Context*. These elements are generated using the LLaMA-3.1-405B model and defined based on the Taxonomy of Culture outlined in (Liu et al., 2024). Our observations suggest that these elements are highly distinguishable and effectively capture key cultural and sociocultural dimensions essential for mapping English idioms to their counterparts in other languages. The prompt used for generating these cultural elements is provided in Appendix A.4.

To construct the **Knowledge Graph (KG)**, we first convert the generated cultural elements into Embeddings (we call it cultural features) with Language-agnostic BERT Sentence Embedding (LaBSE) model (Feng et al., 2022). Once the cul-

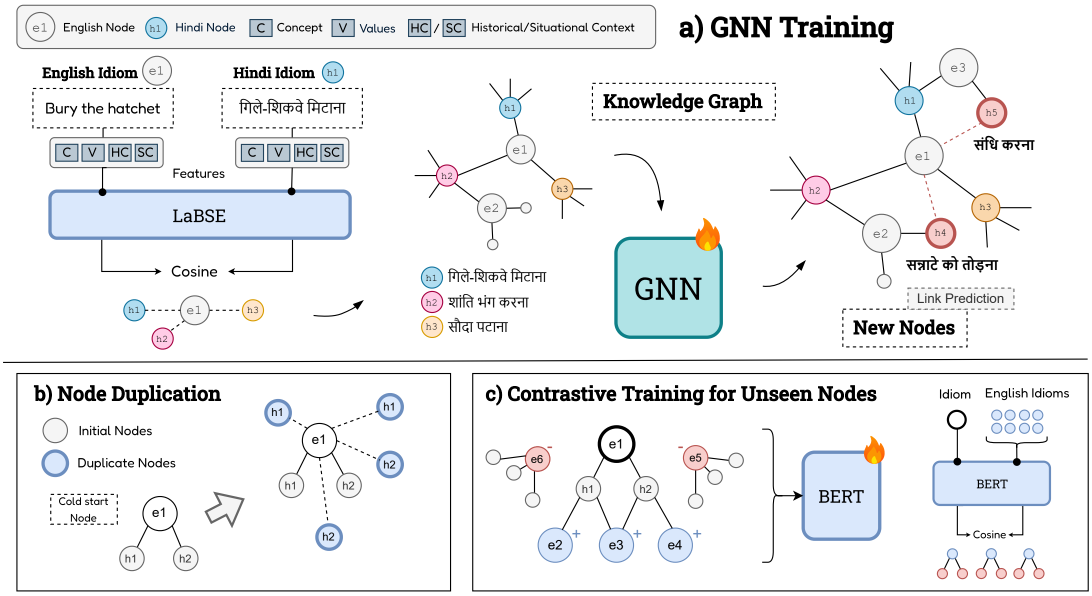

Figure 2: **Overall training process of IdiomCE**: (a) GNN training – illustrating the creation of a Knowledge Graph using source and target idioms, specifically for `en-hi`, leveraging LaBSE embeddings and training a GNN for the Link Prediction (LP) task; (b) Node Duplication – demonstrating how we address the cold start issue by duplicating target nodes; and (c) Contrastive Training – showing the training through positive and negative samples and the process of mapping unseen nodes to relevant target idioms.

tural features for each idiom are generated, we compute the cosine similarity between the cultural features of English and target (Indic) language idioms to establish pairwise mappings, as illustrated in Figure 2. Moreover, to identify the most relevant idiom pairs for the KG, we focus on outliers within the cosine similarity scores, as these indicate strong semantic relationships. Outlier detection is performed by calibrating thresholds based on the skewness and kurtosis of the data, leveraging both the Inter-Quartile Range (IQR) and z-score. By carefully selecting thresholds in these approaches, we ensure that only high-similarity idiom pairs are connected, effectively capturing the most significant relationships. This approach, grounded in robust statistical techniques (Chandola et al., 2009), ensures that the graph reflects the most salient semantic connections.

As a result of this process, multiple KGs are constructed, each linking English idioms to idioms in a specific Indic language. Formally, each KG is represented as $\mathcal{G} \equiv (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the feature of each idiom/node and $\mathcal{E}$ represents the edges connecting source and target idioms.

3.3 IdiomCE

The proposed IdiomCE follows the widely used encoder-decoder architecture for GNN-based link prediction (Kipf and Welling, 2016) (Schlichtkrull et al., 2017) (Zhao et al., 2022) where a GNN encoder learns node representations, and a decoder predicts link existence probabilities for each node pair. Below, we provide a detailed discussion of the training process for our method.

Node Duplication Augmentation. Once the above KG is constructed, we could encounter the cold start problem due to the sparsity of the dataset, which consists of only a few thousand idioms. This issue arises when certain nodes have few or no connections, leading to under-representation in the GNN during the downstream tasks (Hao et al., 2020; Zhang et al., 2023). To mitigate this, we employ a Node Duplication strategy (Guo et al., 2024), which enhances node connectivity and improves representation learning.

We provide a detailed explanation of our node duplication procedure. Let $\mathcal{S}$ and $\mathcal{T}$ represent the sets of source and target language idioms, respectively. For any node $v \in \mathcal{V} \equiv \mathcal{S} \cup \mathcal{T}$, we define its set of neighbors as:

$$\mathcal{N}_v := \{\bar{v} : e_{v,\bar{v}} \text{ or } e_{\bar{v},v} \in \mathcal{N}_v\},$$

where $\mathcal{N}_v$ consists of all nodes $\bar{v}$ connected to v by an edge. We extend the methodology of (Guo et al., 2024) by categorizing source nodes into two types: *Cold nodes* ($\mathcal{T}_{cold}$): Target nodes with fewer than δ neighbors.

Warm nodes ($\mathcal{T}_{warm}$): Target nodes with at least δ neighbors. For our experiment we consider δ equals 3

For each cold node v, we duplicate its neighbors $\mathcal{N}_v$ and create new corresponding source nodes. We then insert edges from v to these duplicated source nodes, as illustrated in Figure 2. In this way, we obtain an augmented graph $\mathcal{G}'$ with these newly created nodes and edges added to the original graph. This approach differs from (Guo et al., 2024), where the authors duplicate source nodes directly based on their degree. In contrast, we duplicate source nodes based on the degree of their corresponding target nodes. This strategy enhances the sampling of under-represented cold nodes by leveraging their connections to source nodes.

IdiomCE Encoder. As discussed in the previous section, once our augmented $\mathcal{G}'$ is created, we convert $\mathcal{G}'$ into the GNN training format by creating a feature vector of each idiom node with a BERT-based embedding model, i.e., LaBSE (Feng et al., 2022). We then construct an initial bi-directional adjacency matrix of edge indices required for training. To ensure generalization across potentially unseen idioms, we employ an inductive GNN for training, specifically SAGEConv (Hamilton et al., 2018). In SAGEConv, each node updates its representation by aggregating the features of its neighbors. The aggregation is done using a permutation invariant function. In our case, we use the mean aggregator, which computes the average of the feature vectors of a node's neighbors. This ensures that the order of neighbors does not affect the result. For a given node v, let $\mathcal{N}(v)$ represent the set of neighbors and h_u denote the features vectors of node u. The mean aggregator is defined as:

$$\mathbf{h}_{\mathcal{N}(v)} = \frac{1}{|\mathcal{N}(v)|} \sum_{u \in \mathcal{N}(v)} \mathbf{h}_u. \tag{1}$$

Next, the node's updated representation is computed by concatenating its own feature vector $\mathbf{h}_v$ with the aggregated neighbor features and then applying a learnable linear transformation followed by a non-linear activation function as given below:

$$\mathbf{h}'_v = \sigma\left(\mathbf{W} \cdot \text{CONCAT}\big(\mathbf{h}_v, \mathbf{h}_{\mathcal{N}(v)}\big)\right). \tag{2}$$

IdiomCE Decoder. We perform the task of link prediction by pairing our IdiomCE encoder with a Multi-Layer Perceptron (MLP) model as a decoder. Given a source node i with GNN embeddings h_i and target node j with GNN embeddings h_j from the Encoder, we first concatenate their embeddings, then pass it through the MLP layer.

$$z_{ij} = [h_i \,\|\, h_j],$$
$$\widehat{y}_{ij} = \text{MLP}(z_{ij}).$$

Once we obtain the prediction from the MLP layer, we then backpropagate using BCE loss and jointly train the GNN and MLP layer for the Link prediction task defined by the loss function given below:

$$\mathcal{L} = -\frac{1}{N} \sum_{(i,j) \in \mathcal{D}} \left[y_{ij} \log \widehat{y}_{ij} + (1 - y_{ij}) \log (1 - \widehat{y}_{ij}) \right]. \tag{3}$$

3.4 Dealing with Unseen nodes

One of the key properties of inductive GNNs is their ability to generalize to unseen nodes, such as idioms absent from the training set. To incorporate an unseen idiom into a trained GNN, it must be connected to relevant neighbors, allowing the model to compute meaningful node embeddings through message passing. A naïve approach is to add edges by randomly selecting target nodes from the initial dataset. However, this often results in dispersed and suboptimal embeddings due to the lack of semantic coherence in the connections. Therefore, to generate high-quality embeddings for an unseen idiom, it is essential to establish connections with semantically relevant neighbors that closely align with its ideal (gold) translation. Given the one-to-many nature of idioms where a single target idiom may correspond to multiple source idioms conveying the same figurative meaning, it is crucial to connect the unseen node to the most similar target idiom neighbors.

To achieve this, we propose training a BERT-based encoder (denoted as $\mathcal{B}_{CL}(\cdot)$) in a contrastive learning setting (Cohan et al., 2020; Ostendorff et al., 2022). The training process leverages a triplet framework designed to align with the graphical structure of our GNN, i.e., $\langle$*anchor* a, *positive* p, *negative* n $\rangle$ where a denotes the source node representing the idiom in the source language, p denotes the source language nodes that are connected to the anchor (i.e., first-hop neighbors in our KG), and n represents nodes that are disconnected (no path exists) to the anchor, ensuring that they do not share semantic similarity. This triplet construction is used in a contrastive loss $\mathcal{L}_t$ that minimizes the distance between the anchor and its positive examples while maximizing the distance to the negative

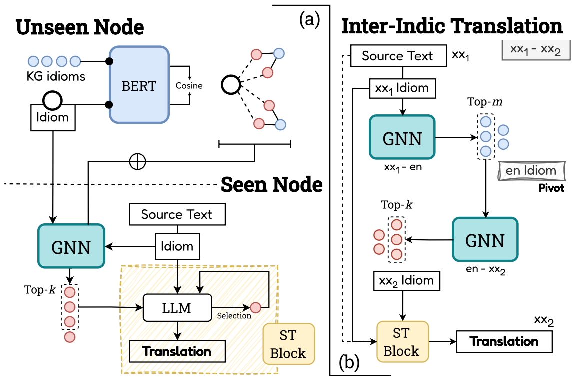

Figure 3: Inference strategy: (a) Unseen & Seen Node Translation – a BERT-trained GNN adapts to unseen nodes, with the Selection and Translation (ST) block selecting idioms via an LLM before translation; (b) Inter-Indic Translation – using English as a pivot between xx_1 and xx_2.

examples. Formally, if h_a, h_p, and h_n are representations of anchor, positive and negative examples, respectively, then with margin α,

$$\mathcal{L}_{\sqcup} = \sum_{(a,p,n)\in\mathcal{D}} \max(0, \|h_a - h_p\| - \|h_a - h_n\| + \alpha).$$

3.5 Inference

From the trained bi-directional GNNs on English and specific Indic languages, we explore idiomatic translation through three approaches, as illustrated in Figure 3: seen nodes, unseen nodes and inter-indic. The *seen nodes*, refer to idioms for which GNN has prior knowledge, including their relationships with other idioms. On the other hand, *unseen nodes* pertain to idioms for which the GNN has no prior information nor any established connections to other idioms. Lastly, *inter-indic* translation where english idioms are treated as pivot, more explanation in section 3.5.3. We assume idiom detection is a well-explored problem, enabling us to focus directly on the translation task without treating idiom identification as an intermediate step. We also presume that the idiom in the source sentence is provided for retrieval through IdiomCE.

3.5.1 Seen Nodes

To infer with *seen nodes*, we first retrieve top-k target idioms using the trained GNN by link prediction by providing source idiom as input. Next, we refine the selection by filtering out the most contextually relevant target idiom based on the source sentence. This is achieved by passing the retrieved idioms into a selection prompt as context in an LLM. Finally, once the most relevant target idiom is identified, we perform LLM-based inference by passing the source text, source idiom, and the selected target idiom into a translation prompt. The details of both prompts are provided in Appendix A.2 and A.3.

3.5.2 Unseen Nodes

For unseen nodes, completely isolated idioms would yield no meaningful results. To address this, we make the following assumption about the training dataset $\mathcal{D}$.

Assumption. For any unseen node u, $\exists v \in \mathcal{D}$ such that $\cos(\mathcal{B}_{CL}(u), \mathcal{B}_{CL}(v)) \geqslant \tau$, where $\tau \in [0, 1]$. For our experiments, we choose τ to be 0.75.

To infer on unseen nodes, we first retrieve the most similar idioms in the source language using cosine similarity based on embeddings from the trained contrastive embedding model $\mathcal{B}_{CL}$. After selecting the top M source language idioms, we randomly select five target-language idioms linked to these source idioms and connect them to the unseen idiom, incorporating them into our graphical data. Once integrated, we perform link prediction on the unseen node to retrieve the most suitable target idiom.

3.5.3 Inter Indic Languages translation

We train the IdiomCE encoder bidirectionally between English and individual Indic languages. In addition to direct translation from $\mathcal{S}$ to $\mathcal{T}$, we propose leveraging trained GNNs for indirect translation. Let $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ be nodes in languages A_1, A_2 and A_3 respectively. Let $\mathcal{G}_{12} : \mathcal{A}_1 \rightarrow \mathcal{A}_2$ and $\mathcal{G}_{23} : \mathcal{A}_2 \rightarrow \mathcal{A}_3$ be GNNs trained between the respective languages. To generate a translation from A_1 to A_3, we use A_2 as the *pivot* language, shown in Figure 3.

4 Experimental set up

Datasets. The initial knowledge graph (KG) construction is based on the dataset from Agrawal et al. (2018) (Agrawal et al., 2018), which provides mappings of idioms between English (en) and seven Indian languages. For our study, we utilize four Indic languages: Tamil (ta), Telugu (te), Bengali (bn), and Hindi (hi). Additionally, we incorporate a parallel idiomatic sentence dataset from Thakre et al. (2018) (Thakre et al., 2018). Beyond these existing resources, we also web-scraped to collect idioms in various Indic languages. For evaluation, we sample 400 sentences from the MAGPIE dataset

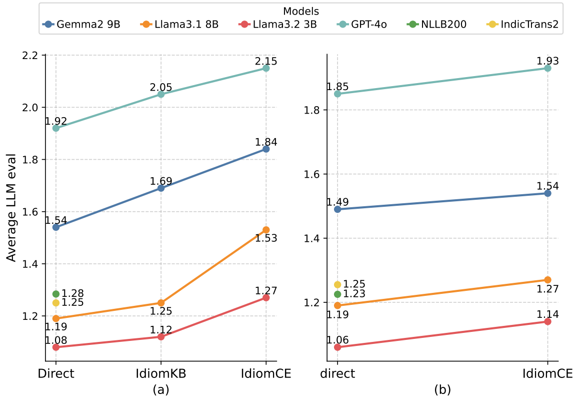

Figure 4: Performance comparison on average LLM score of Models on seen nodes (idiom) (a) and unseen nodes (b) across en-xx direction.

(Haagsma et al., 2020) to assess translation effectiveness from English to Indic languages. To analyze performance under different conditions, we conduct experiments in two setups: (1) Seen Idioms, where idioms present in the training data are tested, and (2) Unseen Idioms, where idioms not encountered during training are evaluated. For the Inter-Indic language setting, we curate a dataset of 200 idiomatic sentences per Indic language from the Samanantar dataset (et al., 2023), ensuring coverage across multiple language pairs, more details on dataset can be found in Appendix A.4.

Evaluation Metrics. Most automatic evaluation metrics, like BLEU (Papineni et al., 2002; Post, 2018) and ChrF (Popović, 2015), rely on reference matching but struggle with one-to-many translation, especially idioms, where n-gram matches fall short. They also fail to distinguish literal from figurative translations. While CometKiwi (Rei, 2022) improves on traditional metrics by being reference-less and semantic-focused, it still struggles to reward high-quality idiomatic translations. Hence, for our evaluation, we adopt the GPT-4o-based evaluation method proposed by (Li et al., 2023a) as our primary metric, as it is an LLM-based approach specifically designed for assessing idiomatic translations we call it here LLM-eval and use WMT22-CometKiwi-DA as a supplementary evaluation metric.

Models. We test the effectiveness of our approach by using base LLMs of varying sizes like Gemma2 9B (Team et al., 2024), Llama-3.1 8B, Llama-3.2 3B (Grattafiori et al., 2024) and GPT-4o mini (OpenAI et al., 2024) in our methodology. We also evaluate our method by comparing them with translations generated from traditional NMT systems like NLLB 3.3B (Team, 2022) and IndicTrans2 (Gala et al., 2023). In our experiments `Direct` represents either directly prompting the LLM to translate the given source sentence, or passing the sentence through the NMT model for generating translation prompt can we referred from Appendix A.1. Specific training details and performance of GNN and MLP layer with other experimental parameters are added in Table 5 in Appendix.

5 Results

Results on Mixed Dataset. This dataset contains a mix of idioms, both seen and unseen during training. We conducted experiments on English-to-Hindi, Bengali, Tamil, and Telugu translation directions. The results in Table 1 show: 1) IdiomCE, our approach that retrieves target idioms based on English idioms, outperforms the direct prompting method, highlighting the effectiveness of our retrieval-based training for idiomatic translation. 2) Among smaller models, Gemma2 9B achieves the best performance, even with direct prompting, demonstrating its strong capabilities in idiomatic translation. 3) With IdiomCE, very small models like Llama 3.2 3B perform comparably to the Directly Prompted larger Llama 3.1 8B variant. 4) Even for larger models like GPT-4o, IdiomCE improves performance, proving its effectiveness across different model sizes. 5) Foundational models like NLLB and IndicTrans2 struggle with idiomatic translation, showing low scores in LLM-eval. On average, IdiomCE improves LLM-eval scores by 18.51% for en-hi, 14.71% for en-bn, 6.45% for en-ta, and 10.33% for en-te. We have also provided example translation in Appendix B. Results on Additional baselines such as (Donthi et al., 2025) are included in Appendix 6.

Results on Seen and Unseen Dataset. In Figure 4, we have shown on average LLM evaluation for different models on various methods across languages. Notably, results for the IdiomKB baseline are shown only for the seen dataset, as IdiomKB supports only idioms present in the training set. On average, the Gemma2 9B model demonstrates the best performance among open-source LLMs on both seen and unseen datasets. Compared to IdiomKB and Direct Method, our approach, IdiomCE, outperform them by 14.28% and 21.78%, respectively, across open-source LLMs for seen dataset. Similarly, for unseen dataset, IdiomCE

Model	Methods	en-hi		en-bn		en-ta		en-te	
		LLM-eval	**COMET**	**LLM-eval**	**COMET**	**LLM-eval**	**COMET**	**LLM-eval**	**COMET**
NLLB-200	Direct	1.3	0.70	1.43	0.769	1.18	0.691	1.1	0.643
Indictrans2	Direct	1.247	0.74	1.275	0.77	1.243	0.769	1.24	0.747
LLama-3.2-3B	IdiomCE	1.34	0.59	1.2	0.6	1.105	0.51	1.18	0.51
	Direct	1.12	0.62	1.05	0.6	1.04	0.52	1.07	0.52
Gemma2-9b-it	IdiomCE	**1.88**	0.68	**1.7**	0.68	**1.63**	0.67	**1.56**	0.62
	Direct	1.6	**0.73**	1.44	**0.71**	1.56	**0.71**	1.46	**0.67**
LLama-3.1-8B	IdiomCE	1.655	0.63	1.40	0.63	1.25	0.57	1.3	0.54
	Direct	1.27	0.68	1.23	0.67	1.16	0.62	1.12	0.59
GPT-4o	IdiomCE	2.39	0.70	2.25	0.69	1.87	0.67	1.83	0.66
	Direct	2.14	0.73	1.99	0.764	1.741	0.72	1.67	0.71

Table 1: Performance Metrics of Various Models on Mixed Dataset; COMET range [0,1].

Model	Methods	hi-xx		bn-xx		ta-xx		te-xx	
		LLM-eval	**COMET**	**LLM-eval**	**COMET**	**LLM-eval**	**COMET**	**LLM-eval**	**COMET**
NLLB-200	Direct	1.85	0.79	1.70	0.78	1.84	0.77	1.81	0.78
Indictrans2	Direct	**1.92**	0.81	1.78	0.81	**2.01**	0.77	1.97	0.77
LLama-3.2-3B	IdiomCE	1.263	0.5663	1.23	0.5867	1.2567	0.53867	1.273	0.5493
	Direct	1.1867	0.589	1.17	0.6163	1.253	0.572	1.1867	0.609
Gemma2-9b-it	IdiomCE	**1.8233**	0.7283	**1.783**	0.727	**1.9867**	0.7267	**2.02**	0.724
	Direct	1.4833	**0.75**	1.49	**0.775**	1.563	**0.755**	1.5467	**0.773**
LLama-3.1-8B	IdiomCE	1.42	0.616	1.46	0.6404	1.533	0.5993	1.493	0.626
	Direct	1.34	0.6533	1.367	0.688	1.25	0.6393	1.25	0.677

Table 2: Performance Metrics of Various Models For Inter-Indic languages; COMET range [0,1].

achieves 5.67% improvement over direct method. Even with GPT-4o results, our approach shows significant improvements for both seen and unseen datasets. Further details on language-specific performance can be found in the Appendix in Table 6 and 7.

Results on Inter-Indic Languages. Table 2 presents the average performance across Indic languages. Our findings indicate: 1)Using English as a pivot to retrieve idioms for translation between Indic languages improves LLM performance compared to direct prompting, highlighting the flexibility of our approach. 2) Gemma2 9B consistently performs well in inter-Indic translation settings, significantly outperforming other LLMs. 3) Interestingly, in some language pairs like hi-xx and ta-xx, IndicTrans2 achieves strong results, even surpassing other models. Overall, IdiomCE demonstrates significant improvements in LLM evaluation, with a 12.5% performance gain for hi-xx, 11.2% for bn-xx, 17.5% for ta-xx, and 19.9% for te-xx translations over Direct prompting.

Methods	en-hi	en-bn	en-tl
IdiomCE	**3.51**	**3.17**	**2.43**
IdiomKB	2.65	1.82	1.88
Direct	2.05	1.58	1.45

Table 3: Human Evaluation on Idiomatic Translation on different methods.

IdiomCE performance under Human evaluation. To compare the performance of IdiomCE with existing baselines, we conducted a manual quality annotation of translations generated by IdiomCE, IdiomKB, and direct translations from Gemma2-9b-it, as this model demonstrated superior performance across the evaluated methods (see Table 1). The evaluation involved 19 native speakers who are highly fluent and bilingual. Assessments were carried out across three language pairs: English–Hindi (en-hi), English–Bengali (en-bn), and English–Telugu (en-tl). Each evaluator was presented with a source sentence containing idiomatic expressions and three corresponding translations produced by the different systems. Evaluators rated each translation on a 5-point scale, with detailed scoring criteria provided in the Appendix 6. As shown in Table A.4, IdiomCE consistently outperformed the other baselines across all three language pairs. The performance gap was especially pronounced in the en-hi and en-bn directions, suggesting that the model is more effective at leveraging GNN-retrieved context for Hindi and Bengali than for Telugu.

Error Analysis. In addition to the human evaluation, we performed an error analysis to identify potential areas for improvement in our methodology. Upon examining the translations, we categorized the observed errors into three distinct types, as outlined below:

- **Morphological Issues.** In some cases, Llama 3.1 8B and Llama 3.2 3B directly replaced an idiom without adapting its morphology, leading to unnatural phrasing in the target language. This suggests that smaller models struggle with idiom adaptation, whereas larger models perform better by adjusting idiomatic structures to fit grammatical norms. These observations highlight scalability challenges in idiomatic translation for smaller models, emphasizing the need for additional fine-tuning or external knowledge integration for improved performance.

- **Incorrect Selection.** In smaller models, such as LLaMA 3.2 3B, the model struggles to correctly select the appropriate target idiom for translation. This issue persists even when the GNN Top-K retrieval includes high-quality idiomatic translations. We have observed this phenomenon more frequently in languages such as Tamil and Telugu.

- **Pivot Noise.** For inter-Indic translations, we employ English as a pivot language to facilitate translation from one Indic language to another, leveraging the bidirectional property of GNN. However, this approach introduces potential noise, which can result in the best target-language idioms ranking lower in the Top-K retrieval. In some cases, high-quality idiomatic translations are entirely excluded from the retrieved set, leading to inaccuracies in the final translation.

Ablation Studies. To justify the use of the Node Duplication procedure (see Sec 3.2), we conduct an ablation experiment comparing performance with and without the NodeDup module in Table 4. We report Hits@k (Chen et al., 2020) for the en-hi translation task, which includes 8,233 nodes (4.6K Hindi target nodes), with 1.1K cold target nodes. Our results show that incorporating the NodeDup module improves Hits@k by 4.85% for $k = 5$ and 6.97% for $k = 10$, demonstrating its effectiveness in enhancing target node retrieval.

Hits @k	Without NodeDup	With NodeDup
Hits@5	81.33 ± 2.36	**85.28** ± 2.99
Hits@10	90.00 ± 2.36	**96.28**± 1.37
Hits@20	100.00 ± 0.00	100.00 ± 0.00
Hits@50	100.00 ± 0.00	100.00 ± 0.00
AUC	95.32	**96.33**

Table 4: Ablation on Node Duplication module.

6 Conclusion

In this work, we addressed the challenges of idiomatic translation by introducing IdiomCE, an adaptive GNN-based approach that effectively captures the complex relationships between idiomatic expressions across languages. Our method generalizes to seen and unseen idioms, improves translation quality even in smaller models, and enables translation via a pivot language through the GNN framework. Experimental results across multiple Indian languages demonstrate that our approach outperforms traditional static knowledge graphs and prompt-based methods, significantly improving idiomatic translation. By leveraging GPT-4 as an evaluation metric, we show that our model better preserves meaning and cultural nuances in translation. Future work can extend this approach to more languages and richer contextual signals.

Limitations

While our work shows significant improvements in idiomatic translation, we mention some of the limitations of our work. Our approach heavily depends on the synthetically generated cultural elements (features). Noisy features, especially in low-resource languages, might affect the performance of our method. Secondly, as mentioned before, although our model captures idiomatic mappings, some idioms rely heavily on a deep contextual understanding of the surrounding sentences and not just on the training data used, which can limit the model's performance.

References

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, and Dipti Sharma. 2018. No more beating about the bush : A step towards idiom handling for Indian language NLP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700, Dubrovnik, Croatia. Association for Computational Linguistics.

Antonio Castaldo and Johanna Monti. 2024. Prompting large language models for idiomatic translation. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 32–39, Sheffield, United Kingdom. European Association for Machine Translation.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3).

Zhe Chen, Yuehan Wang, Bin Zhao, Jing Cheng, Xin Zhao, and Zongtao Duan. 2020. Knowledge graph completion: A review. *IEEE Access*, 8:192435–192456.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM abilities in idiomatic translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gowtham Ramesh et al. 2023. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Preprint*, arXiv:2104.05596.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, 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. *Preprint*, arXiv:2305.16307.

Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhichun Guo, Tong Zhao, Yozen Liu, Kaiwen Dong, William Shiao, Neil Shah, and Nitesh V. Chawla. 2024. Node duplication improves cold-start link prediction. *Preprint*, arXiv:2402.09711.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2018. Inductive representation learning on large graphs. *Preprint*, arXiv:1706.02216.

Bowen Hao, Jing Zhang, Hongzhi Yin, Cuiping Li, and Hong Chen. 2020. Pre-training graph neural networks for cold-start users and items representation. *Preprint*, arXiv:2012.07064.

Kazuma Hashimoto and Yoshimasa Tsuruoka. 2016. Adaptive joint learning of compositional and non-compositional phrase embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 205–215, Berlin, Germany. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2016. Variational graph auto-encoders. *Preprint*, arXiv:1611.07308.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2023a. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. *Preprint*, arXiv:2308.13961.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2023b. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. *Preprint*, arXiv:2308.13961.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Preprint*, arXiv:2406.03930.

Jatin C. Modh and Jatinderkumar R. Saini. 2020. Context based mts for translating gujarati trigram and bigram idioms to english. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–6.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. In Stephen Everson, editor, *Language: Companions to Ancient Thought, Vol. 3*, pages 491–538. Cambridge University Press.

OpenAI, Josh Achiam, Steven Adler, and Sandhini Agarwal et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *Preprint*, arXiv:2411.00860.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. 2023. Knowledge graphs: Opportunities and challenges. *Preprint*, arXiv:2303.13948.

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, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.

Ricardo et al. Rei. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sara Rezaeimanesh, Faezeh Hosseini, and Yadollah Yaghoobzadeh. 2024a. A comparative study of llms, nmt models, and their combination in persian-english idiom translation. *Preprint*, arXiv:2412.09993.

Sara Rezaeimanesh, Faezeh Hosseini, and Yadollah Yaghoobzadeh. 2024b. A comparative study of llms, nmt models, and their combination in persian-english idiom translation. *Preprint*, arXiv:2412.09993.

Giancarlo Salton, Robert Ross, and John Kelleher. 2014. Evaluation of a substitution method for idiom transformation in statistical machine translation. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 38–42, Gothenburg, Sweden. Association for Computational Linguistics.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. Modeling relational data with graph convolutional networks. *Preprint*, arXiv:1703.06103.

Naziya Shaikh. 2020. Determination of idiomatic sentences in paragraphs using statement classification and generalization of grammar rules. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 45–50, Marseille, France. European Language Resources Association (ELRA).

Gemma Team, Morgane Riviere, and Shreya Pathak et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

NLLB Team. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

N. Thakre, V. Gupta, and N. Joshi. 2018. Identification and translation of idiomatic sentence from hindi to english. *International Journal of Computer Sciences and Engineering*, 6(5):283–287.

Elsa Vula and Nazli Tyfekçi. 2024. Navigating non-literal language: The complexities of translating idioms across cultural boundaries. *Academic Journal of Interdisciplinary Studies*, 13:284.

Marion Weller, Fabienne Cap, Stefan Müller, Sabine Schulte im Walde, and Alexander Fraser. 2014. Distinguishing degrees of compositionality in compound splitting for statistical machine translation. In *Proceedings of the First Workshop on Computational Approaches to Compound Analysis (ComAComA 2014)*, pages 81–90.

Oktay Yagiz and Siros Izadpanah. 2013. Language, culture, idioms, and their relationship with the foreign language. *Journal of Language Teaching and Research*, 4.

Andrea Zaninello and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France. European Language Resources Association.

Xiangyu Zhang, Zongqiang Kuang, Zehao Zhang, Fan Huang, and Xianfeng Tan. 2023. Cold & warm net: Addressing cold-start users in recommender systems. *Preprint*, arXiv:2309.15646.

Tong Zhao, Gang Liu, Daheng Wang, Wenhao Yu, and Meng Jiang. 2022. Learning from counterfactual links for link prediction. *Preprint*, arXiv:2106.02172.

A Prompts used in the experiments

A.1 Direct Prompt

> Translate the Following {src_lang} Sentence to {tgt_lang}. Only provide final translation as output, Do not provide any explainations.
>
> **{src_lang} Sentence:** {sent}

A.2 Selection Prompt

> You are a linguistic researcher on idioms and good at {tgt_lang} and {src_lang}. Choose the best {tgt_lang} idiom matching the {src_lang} idiom and Context of Source Sentence in which it is used in. Only Provide Best macthing {tgt_lang} Idiom Do not provide any explaination.
>
> **{src_lang} idiom:** {en_idm}
> **Source Sentence:** {sent}
> **Options:** {tgt_lang idioms}

A.3 Translation Prompt

> You are a linguistic researcher on idioms and are good at {tgt_lang} and {src_lang}. {en_idm} means {hi_idm}. Given the above knowledge, translate the following sentence to {tgt_lang}: **{sent}**.
>
> Only provide final translation as output, Do not provide any Explainations.

A.4 Cultural element generation prompt

```
You are a linguistic expert with deep
knowledge of {tgt_lang} idioms,
including their cultural and socio-
cultural contexts. For the given
idiom, provide a detailed analysis
covering the following key aspects.
Ensure each point has only a brief,
single-sentence description:
1. **Idiom:** - {idiom}
2. **Concepts:** - Explain the basic
meaning and underlying concepts of
the idiom.
3. **Values:** - Describe the beliefs
or desirable outcomes that the idiom
reflects.
4. **Situational Context:** -
Describe typical scenarios where the
idiom is used.
5. **Historical Context:** - Provide
any relevant historical background
influencing the idiom's usage.
```

Training Details. We train the GNN using a 2-layer SAGEConv architecture, mapping input states from 768 to a hidden dimension of 64. The hidden representation then passes through an MLP with two linear layers and ReLU activation. The model is trained for 50 epochs over 5 runs. For Node Duplication Augmentation, each target node is duplicated twice, and the distinction threshold (δ) between cold and warm nodes is set to 3. α used as margin for Contrastive Training is set to 1.

Human Evaluation Instruction. Here we provide details on Human Evaluation Instruction, we conducted human evaluation on three translation directions: English to Hindi, English to Bengali, and English to Telugu. The evaluation compared three methods: IdiomCE (ours), IdiomKB, and Direct Prompting results are shared in 3. A total of 19 native speakers, who are highly fluent and bilingual, participated in the evaluation. Each evaluator was presented with: 1) A source sentence containing an idiom. 2) Three translations generated by the different methods. Instructions to score the translations on a scale of 1 to 5, based on the following criteria:

- **Score 1:** The sentence is correctly translated, but the idiom is completely mistranslated, missing its figurative meaning or translated literally.
- **Score 2:** The sentence is correctly translated, and the idiom is translated, but it does not fully convey the intended meaning.
- **Score 3:** The sentence and idiom are correctly translated, but the idiomatic expression does not sound natural to native speakers.
- **Score 4:** The sentence and idiom are accurately translated, highly natural, and the overall translation is fluent.
- **Score 5:** The translation is perfectly natural for native speakers, with the idiom translated in the best possible way. This evaluation provides insights into how well each method captures idiomatic expressions while maintaining fluency and naturalness.

Dataset Composition. The total number of unique idioms per language in our training dataset is as follows: Telugu - 4,407, Bengali - 4,479, Tamil - 4,179, Hindi - 4,722, and English – 4,500. When this data is transformed into a graphical structure, the training dataset—prior to applying the Node Duplication Augmentation strategy the Training Composition for GNN expands to:

- English and Tamil: 7,646
- English and Telugu: 7,988
- English and Bengali: 7,872
- English and Hindi: 8,233

The test set, as detailed in Section 4 (Datasets), it consists of 400 unique idiom-containing sentences, with 200 sentences featuring seen idioms and 200 sentences containing unseen idioms. This is referred to as the Mixed Dataset in our paper. Additionally, for the inter-Indic translation setting, we include 200 sentences, each containing a unique idiom.

Computational Resource and Inference time. We thank the reviewer for their suggestion. Below is a detailed breakdown of the computational resources used: GPU Specifications: The experiments were conducted on an NVIDIA GeForce RTX 3090 with CUDA version 12.2 and 24GB of VRAM. Training Time: Training the GNN for a single language direction takes approximately 6–7 minutes, with a total GPU memory requirement of around 500MB. Inference Time and GPU utilization:

- On the seen dataset (200 sentences): 14–15 minutes.
- On the unseen dataset (200 sentences): 20–22 minutes.
- For the inter-Indic translation setting (200 sentences): 17–18 minutes.

Overall GPU Utilization: When the entire inference pipeline runs simultaneously, the maximum GPU memory required is approximately 10–11GB.

B Examples

B.1 Translation Example en-hi direction

Source: **Close call**, though.

Direct: करीब से निकल गया।

IdiomCE: पीछा छुड़ा, हालाँकि...

Source: Well, **it's all Greek to me**, but so long as it keeps him happy.

Direct: अच्छा, यह मेरे लिए सब ग्रीक है, लेकिन जब तक यह उसे खुश रखता है।

IdiomCE: चाहे वो मुझे समझ न आए, पर जब तक उसे खुश रखता है।

Source: Other women took one look at them, and went **green with envy**.

Direct: दूसरी महिलाओं ने उन्हें एक झलक में देखा और ईर्ष्या से हरा रंग हो गया।

IdiomCE: दूसरी महिलाओं ने उन्हें एक झलक में देखा और जल भून गईं।

Source: You're **chasing** an impossible **rainbow**.

Direct: आप असंभव इंद्रधनुष का पीछा कर रहे हैं।

IdiomCE: तुम खयाली पुलाव पका रहे हो।

Source: And for God's sake **bury the hatchet**.

Direct: और भगवान की आज्ञा से, तलवारों को दफना दो।

IdiomCE: और भगवान के लिए शत्रुता खत्म कर दो।

B.2 Translation Example en-bn direction

Source: A bottom drain is something which has **stood the test of time** too.

Direct: একটি বটম ড্রেনও বহুকাল ধরিয়া বিদ্যমান।

IdiomCE: একটি বটম ড্রেনও সময়ের প্রতিবাদে স্থিতিশীল।

Source: I hope you will soon **find your feet** again.

Direct: আমি আশা করি তুমি শীঘ্রই পা পাতা নেওয়া।

IdiomCE: আমি আশা করি, তুমি শীঘ্রই আবার তোমার আধার পেয়ে যাবে।

Source: We are not **joined at the hip**, Cabochon said, and smiled bitterly.

Direct: আমরা হাতের কাছে হাত বাড়িয়ে নেওয়া নই, ক্যাবোচন বলেছিল, এবং ক্ষুব্ধ হাসি দিয়ে।

IdiomCE: আমরা হিপে যুক্ত নই, ক্যাবোচন বলেছিল, এবং ক্ষুব্ধ হাসি দিয়ে।

Source: After all, **charity begins at home**.

Direct: অবশ্যই, দান নিজের ঘরের দেখা তারপর পরের ঘর।

IdiomCE: অবশেষে, দান বাড়ির থেকে শুরু হয়।

Source: Youd both have **got on like a house on fire**.

Direct: আপনারা উভয় খুব বাড় বাড়া হয়েছিল।

IdiomCE: আপনি দুজনেই খুব ভালো মিলেত।

B.3 Translation Example en-ta direction

Source: I had to **face the music**, I had to face myself.

Direct: நான் இசையை எதிர்கொள்ள வேண்டியிருந்தது, நான் என்னைத் தானே எதிர்கொள்ள வேண்டியிருந்தது.

IdiomCE: நான் படாத பாடு படுவதற்கு ஆக வேண்டியிருந்தது, நான் என்னை சந்திக்க வேண்டியிருந்தது.

Source: If I had to **sweat blood** it would be done.

Direct: எனக்கு இரத்தம் சிந்தினாலும், இது முடிந்துவிடும்.

IdiomCE: உயிரைக் கொடுத்து உழைத்தாலும் முடித்துவிடுவேன்.

Source: As Crilly is taken back into the cells, he **catches my eyes**.

Direct: கிரிலி செல் மீண்டும் கொண்டு செல்லப்படும் போது, என் கண்களைப் பிடித்துக்கொள்கிறார்.

IdiomCE: கிரில்லி மீண்டும் சிறைச்சாட்டிற்கு அழைத்துச் செல்லும் போது, அவர் என் கவனத்தை ஈர்க்குகிறார்.

Source: Why should he suddenly have materialized **out of the blue**?

Direct: அவன் திடீரென்று எப்படி மறைந்து போனான்?

IdiomCE: அவன் திடீரென்று எப்படி தோன்ற முடியும்?

Source: In a nutshell Yes.

Direct: சொல்லப்போனால், ஆம்.

IdiomCE: சுருக்கமாக சொல்வதென்றால் ஆம்.

B.4 Translation Example en-te direction

Source: You just have to try to **keep your head above water**.

Direct: నీవు కేవలం నీ తల నీటిపై ఉంచుకోవడానికి ప్రయత్నించాలి.

IdiomCE: తావు తీవ్రంగా ప్రయత్నం చేయాలి.

Source: In every country, intellectuals, too, have **jumped on the nationalist bandwagon**.

Direct: ప్రతి దేశంలో, మాకు

IdiomCE: ప్రతి దేశంలో, విద్వాంసులు కూడా జాతీయవాద బండిపై గోల గోల చేస్తున్నారు.

Source: **Keep your chin up**, girl, were not lost yet.

Direct: ముఖం ఎత్తు, అమ్మాయి, మాకు ఇంకా కోల్పోయలేదు.

IdiomCE: తల ఎత్తుకుని ఉండు, అమ్మాయి, మాకు ఇంకా కోల్పోయలేదా.

Source: Poor old British Rail were **between the devil and the deep blue sea**.

Direct: బాధపడిన పాత బ్రిటిష్ రైల్ దెవదాంబుల మధ్య ఉంది.

IdiomCE: గట్టిగా పోరాడిన బ్రిటిష్ రైల్ ముందు నుయ్యి వెనక గొయ్యిలో ఉన్నారు.

Source: **Close, but no cigar**.

Direct: సమీపంలో ఉన్నా, కానీ సిగారు కాదు.

IdiomCE: దగ్గరకు వచ్చినా దక్కలేదు.

Language	Hits@5	Hits@10	Hits@20	Hits@50	AUC
hindi	85.28 ± 2.99	96.28 ± 1.37	100.00 ± 0.00	100.00 ± 0.00	96.33 ± 0.28
Telugu	82.50 ± 8.54	95.83 ± 2.95	100.00 ± 0.00	100.00 ± 0.00	95.32 ± 0.37
Tamil	76.06 ± 3.98	88.45 ± 2.09	98.59 ± 1.00	100.00 ± 0.00	93.27 ± 0.73
Bengali	79.29 ± 5.30	95.00 ± 4.07	99.29 ± 1.60	100.00 ± 0.00	96.10 ± 0.12

Table 5: Performance on GNN Link Prediction task for each language.

Model	**Methods**	en-hi		en-bn		en-ta		en-te	
		GPT-4	**COMET**	**GPT-4**	**COMET**	**GPT-4**	**COMET**	**GPT-4**	**COMET**
NLLB-200	Direct	1.34	0.70	1.45	0.77	1.21	0.69	1.14	0.64
Indictrans2	Direct	1.24	0.74	1.27	0.78	1.26	0.76	1.21	0.74
LLama-3.2-3B	IdiomCE	1.42	0.58	1.26	0.59	1.15	0.52	1.24	0.51
	Direct	1.12	0.62	1.06	0.60	1.03	0.51	1.09	0.54
	IdiomKB	1.25	0.61	1.05	0.59	1.07	0.52	1.11	0.52
	LIA	1.13	0.565	1.01	0.5702	0.97	0.510	1.06	0.491
	SIA	1.18	0.57	1.15	0.58	1.10	0.53	1.09	0.48
Gemma2-9b-it	IdiomCE	2.08	0.69	1.84	0.69	1.76	0.68	1.68	0.63
	Direct	1.63	0.73	1.50	0.71	1.60	0.72	1.45	0.68
	IdiomKB	1.875	0.70	1.64	0.70	1.65	0.68	1.50	0.64
	LIA	1.30	0.64	1.18	0.610	1.184	0.554	1.125	0.517
	SIA	1.41	0.65	1.30	0.63	1.23	0.562	1.20	0.55
LLama-3.1-8B	IdiomCE	1.89	0.62	1.54	0.63	1.29	0.57	1.41	0.53
	Direct	1.27	0.68	1.22	0.67	1.16	0.62	1.14	0.58
	IdiomKB	1.40	0.67	1.19	0.67	1.20	0.60	1.21	0.59
	LIA	1.60	0.645	1.50	0.632	1.42	0.61	1.35	0.56
	SIA	2.08	0.69	1.84	0.69	1.60	0.65	1.68	0.63

Table 6: Performance Metrics of Various Models on Seen Dataset; COMET range [0,1].

Model	**Methods**	en-hi		en-bn		en-ta		en-te	
		LLM-eval	**COMET**	**LLM-eval**	**COMET**	**LLM-eval**	**COMET**	**LLM-eval**	**COMET**
NLLB-200	Direct	1.26	0.70	1.41	0.77	1.17	0.69	1.06	0.64
Indictrans2	Direct	1.25	0.74	1.28	0.78	1.22	0.76	1.27	0.74
LLama-3.2-3B	IdiomCE	1.25	0.58	1.14	0.59	1.06	0.52	1.13	0.51
	LIA	1.13	0.564	1.05	0.570	1.06	0.508	1.02	0.491
	Direct	1.12	0.62	1.05	0.60	1.05	0.51	1.05	0.53
Gemma2-9b-it	IdiomCE	1.68	0.68	1.56	0.67	1.50	0.68	1.49	0.66
	LIA	1.34	0.62	1.17	0.610	1.11	0.56	1.12	0.517
	Direct	1.57	0.72	1.39	0.70	1.53	0.72	1.4	0.68
LLama-3.1-8B	IdiomCE	1.42	0.63	1.27	0.62	1.21	0.59	1.19	0.53
	LIA	1.61	0.65	1.50	0.653	1.40	0.64	1.38	0.587
	Direct	1.28	0.68	1.23	0.67	1.16	0.62	1.11	0.55

Table 7: Performance Metrics of Various Models on Unseen Dataset; COMET range [0,1].