

# BanglaAutoKG: Automatic Bangla Knowledge Graph Construction with Semantic Neural Graph Filtering

Azmine Toushik Wasi <sup>1</sup>, Taki Hasan Rafi <sup>2</sup>, Raima Islam <sup>3</sup>, Dong-Kyu Chae <sup>2</sup>

<sup>1</sup>Shahjalal University of Science and Technology, Bangladesh <sup>2</sup>Hanyang University,  
<sup>3</sup>BRAC University, Bangladesh

LREC-COLING  2024

(Short Paper)

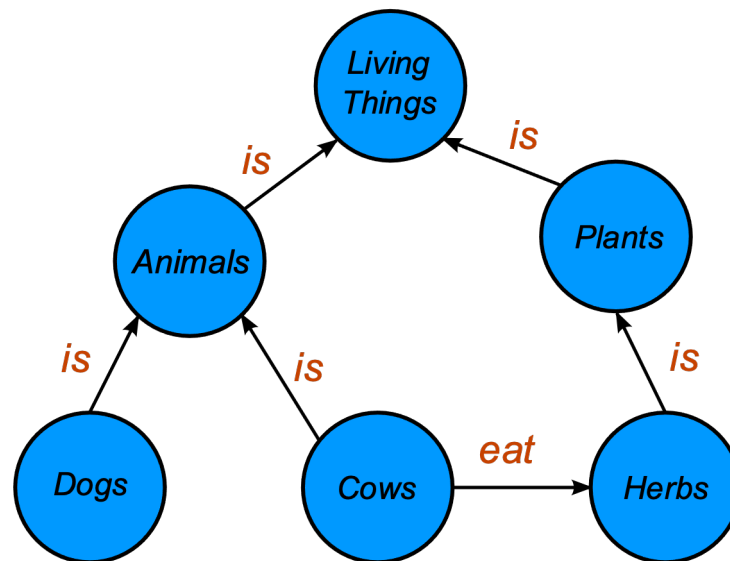


# 1. Background

---

## What is Knowledge Graph?

- Knowledge Graphs (KGs) are semantic graphs consisting of large collections of factual entities and relations, which depict knowledge of real-world objects.



## 2. Motivation

---

However, there is no Knowledge Graph (KG) system available for **Bangla** language.

### Reasons:

- ✓ No KG research
- ✓ Lack of comprehensive resources
  - No proper datasets
  - No universal language models like GPT-3 or BERT
- ✓ No universal NER models



# Contributions

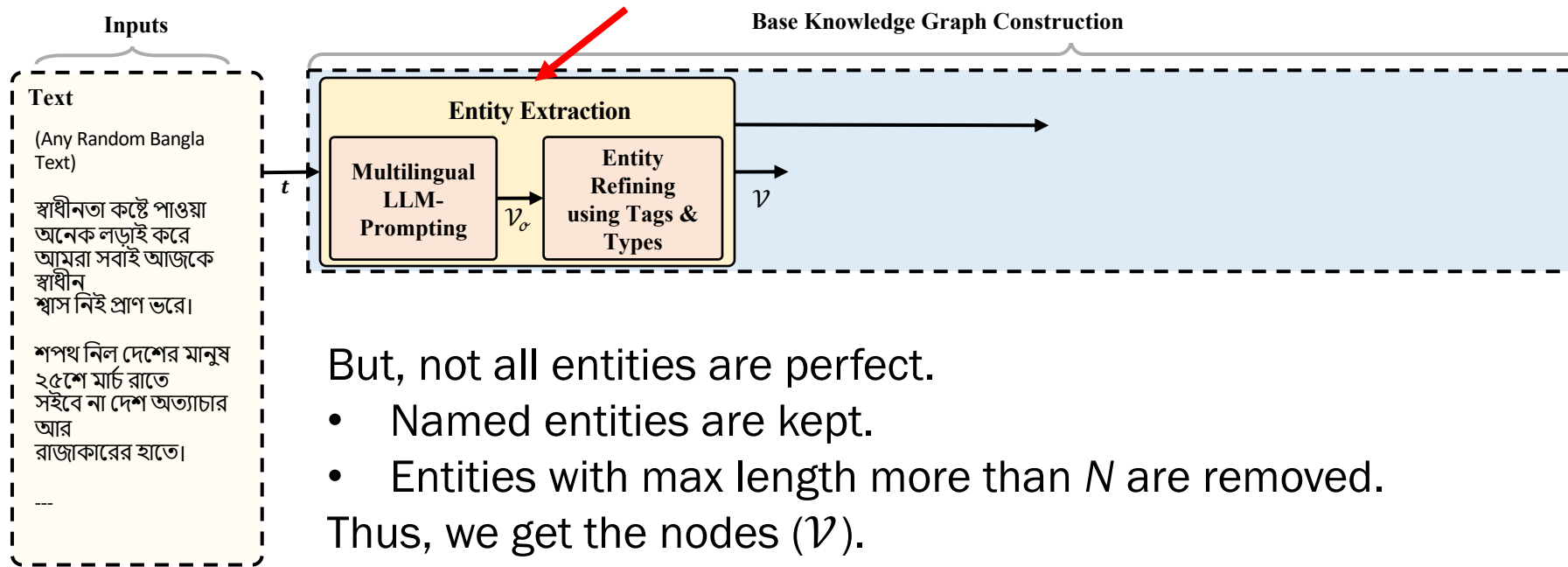
---

- We develop a novel and universal automatic KG generation framework with semantic-filtered for Bangla language.
- We construct a universal KG by utilizing multilingual LLMs, pre-trained BERT based feature development and alignment within entities. We also develop a semantic filtering method to streamline KGs by removing unnecessary edges.
- Our model is able to construct Bengali KGs from texts effectively.



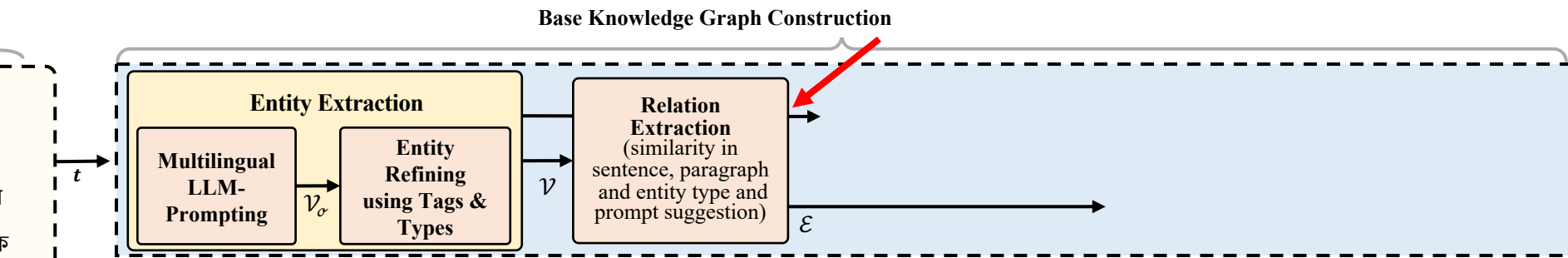
# 3. Approach

To build the base KG from any given random Bangla text  $t$ , at first, we prompt in a multilingual LLM to find raw entities ( $\mathcal{V}_\sigma$ ).



# 3. Approach

For constructing edges ( $\mathcal{E}$ ),



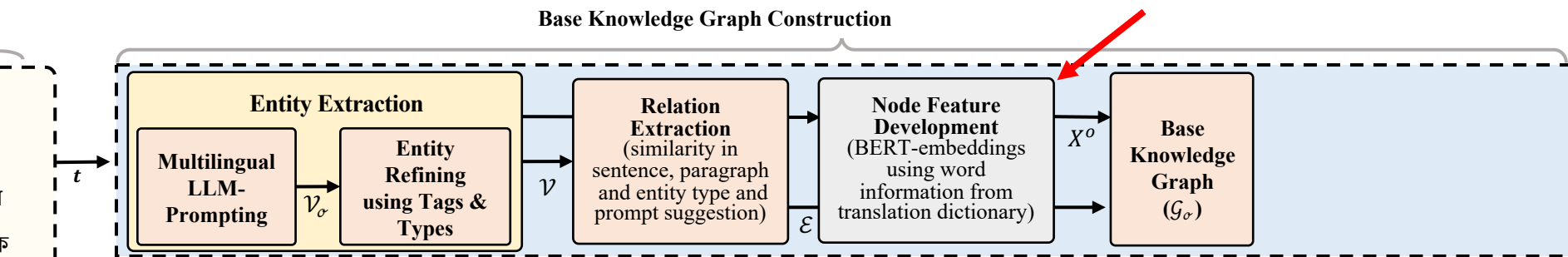
We extract relations of the entities by

- Similarities in sentence and paragraph
- Similarity between entity types
- LLM suggestions



# 3. Approach

For getting node features ( $X^o$ ),



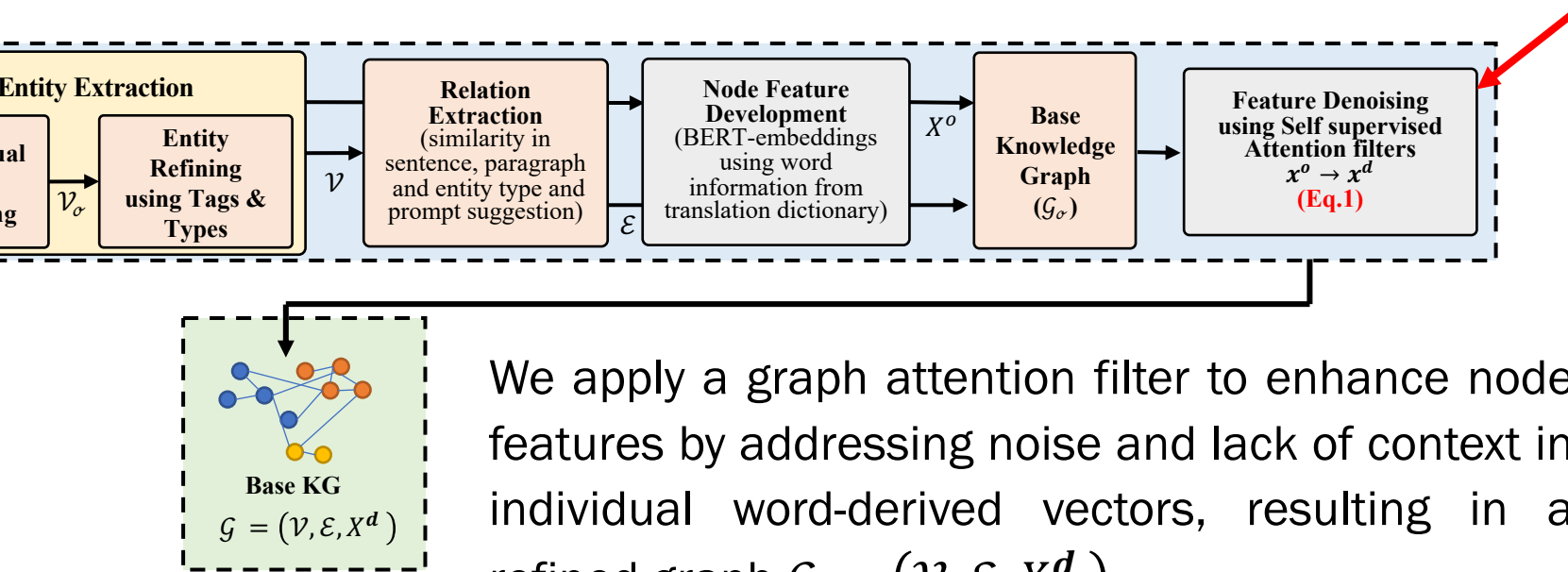
We extract BERT embeddings of the entities from their English translation using a translation dictionary and LLM output suggestion.

Combining  $\mathcal{V}$ ,  $\mathcal{E}$  and  $X^o$ , we construct the base knowledge graph,  $\mathcal{G}_\sigma$



# 3. Approach

But, as these node features are obtained from BERT, they have high noise and almost no correlation.



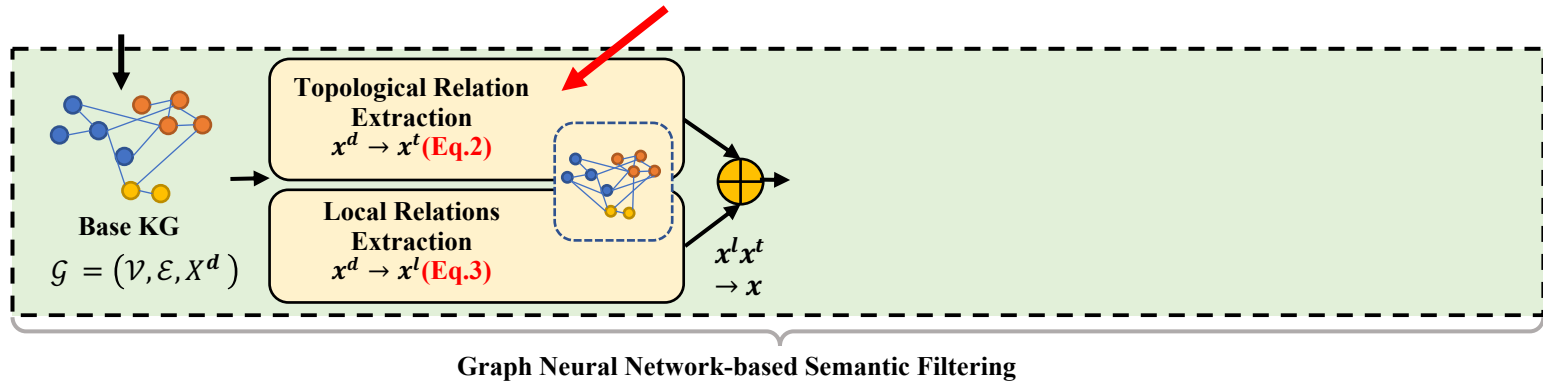
We apply a graph attention filter to enhance node features by addressing noise and lack of context in individual word-derived vectors, resulting in a refined graph  $G = (v, \epsilon, X^d)$ .

$$x_i^d = \alpha_{ii} W_{FD} x_i^o + \sum_{j \in \mathcal{N}(i)} \alpha_{ij} W_{FD} x_j^o$$



# 3. Approach

As our Base KG is ready, it's time for **semantic filtering**.



First step of semantic filtering is **Information Extraction**. It involves re-weighting edges based on topological and local information through attention-based graph convolutions.

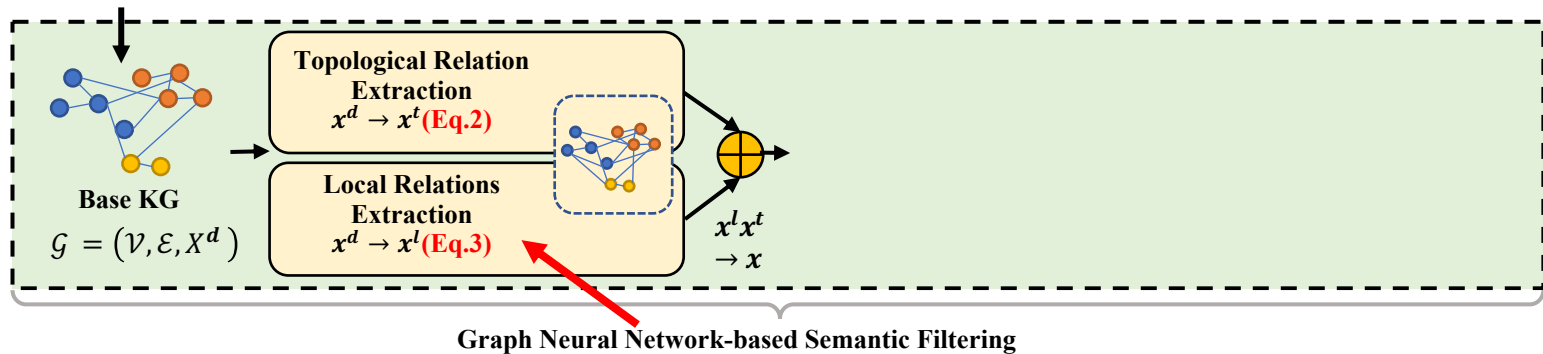
- **Topological Relation Extraction:** Applies attention-based conv. to transform node features for topological relations within  $\mathcal{G}$  ( $x^d \rightarrow x^t$ ).

$$x_i^t = W_{TR}^T \sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{e_{j,i}}{\sqrt{\hat{d}_j \hat{d}_i}} x_j^d$$



# 3. Approach

Then comes Local Relation Extraction.



- **Local Relation Extraction:** Utilizes spectral filtering to extract local neighborhood relations, enhancing KG construction through recursive feature transformations ( $x^d \rightarrow x^l$ ).

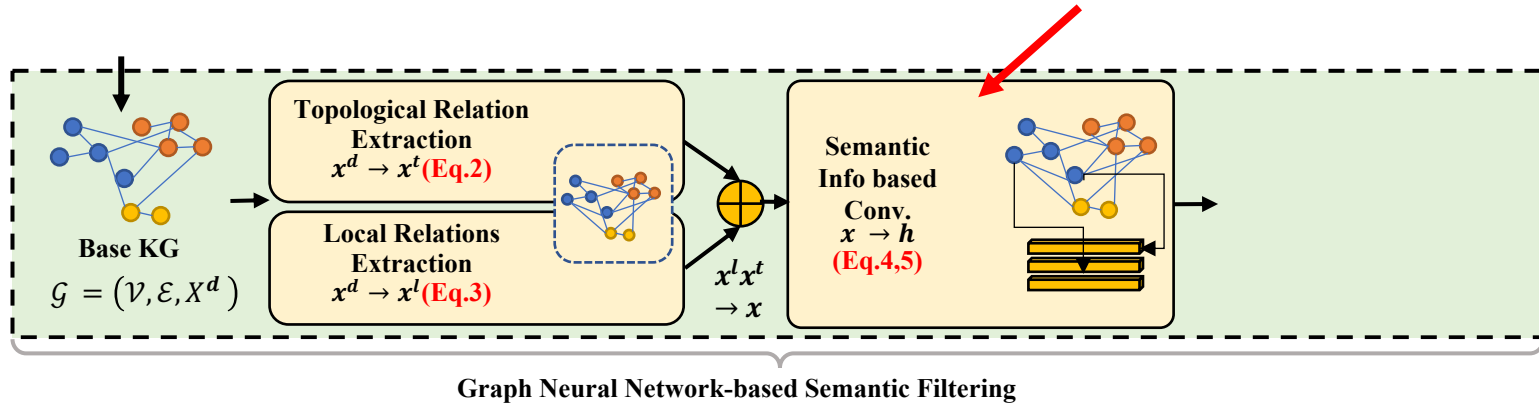
$$X^l = \sum_{k=1}^K Z^{(k)} \cdot W_{NR}^{(k)}$$

Then we combine both topological and local feature representations to form a unified feature matrix  $X$  ( $X = [X^t, X^l]$ ),



# 3. Approach

Then comes Semantic Information Convolution.



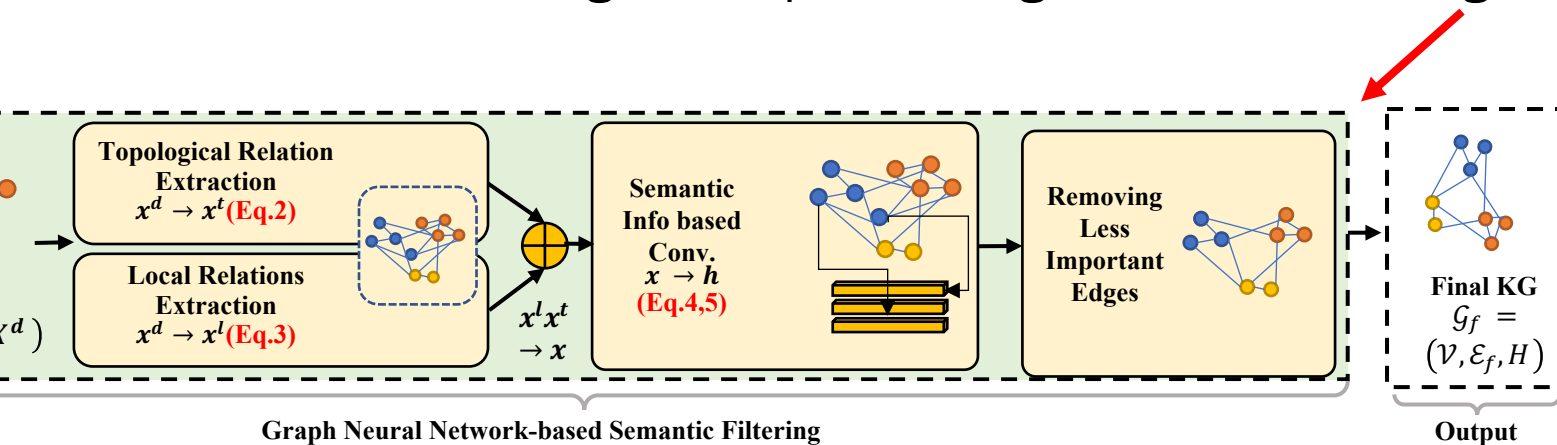
We apply attention-based convolution for final node feature extraction.

$$h_i = \alpha W_S x_i + \sum_{j \in \mathcal{N}(i)} \alpha_{ij} W_S x_j$$



### 3. Approach

Then comes removing less important edges and constructing the final KG.



We use semantic feature similarity to eliminate redundant or less important edge, resulting in the final KG,

$$G_f = (V, E_f, H).$$



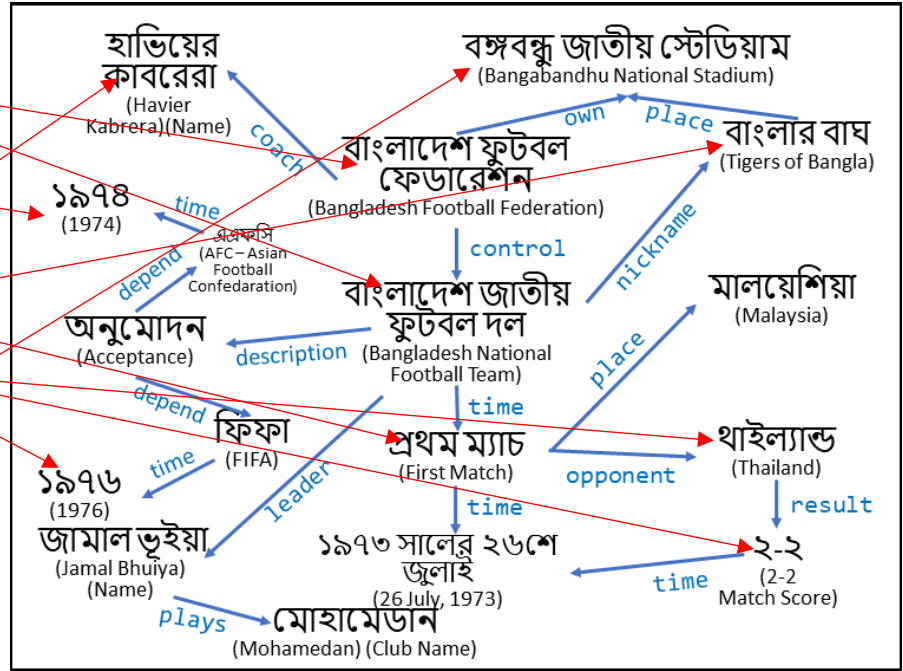


# 4. Case Studies

## Wikipedia article:

বাংলাদেশ জাতীয় ফুটবল দল হচ্ছে আন্তর্জাতিক ফুটবলে বাংলাদেশের প্রতিনিধিত্বকারী পুরুষদের জাতীয় দল, যার সকল কার্যক্রম বাংলাদেশের ফুটবলের সর্বোচ্চ নিয়ন্ত্রক সংস্থা বাংলাদেশ ফুটবল ফেডারেশন দ্বারা নিয়ন্ত্রিত হয়। এই দলটি ১৯৭৬ সাল হতে ফুটবলের সর্বোচ্চ সংস্থা ফিফার এবং ১৯৭৪ সাল হতে তাদের আঞ্চলিক সংস্থা এশিয়ান ফুটবল কনফেডারেশনের সদস্য হিসেবে রয়েছে। ১৯৭৩ সালের ২৬শে জুলাই তারিখে, বাংলাদেশ প্রথমবারের মতো আন্তর্জাতিক খেলায় অংশগ্রহণ করেছে; মালয়েশিয়ার কুয়ালালামপুরে অনুষ্ঠিত বাংলাদেশ এবং থাইল্যান্ডের মধ্যকার উক্ত ম্যাচটি ২-২ গোলে ড্র হয়েছে।

৩৬,০০০ ধারণক্ষমতাবিশিষ্ট বঙ্গবন্ধু জাতীয় স্টেডিয়ামে বাংলার বাঘ নামে পরিচিত এই দলটি তাদের সকল হোম ম্যাচ আয়োজন করে থাকে। এই দলের প্রধান কার্যালয় বাংলাদেশের রাজধানী ঢাকার মতিঝিলের বঙ্গবন্ধু জাতীয় স্টেডিয়ামের নিকটবর্তী বিএফএফ ভবনে অবস্থিত। বর্তমানে এই দলের ম্যানেজারের দায়িত্ব পালন করছেন হাভিয়ের কাবেরেরা এবং অধিনায়কের দায়িত্ব পালন করছেন কলকাতা মোহামেডানের মধ্যমাঠের খেলোয়াড় জামাল ভূইয়া।



## 6. Discussions and Insights

---

- **Bengali-specific Text Encoders:** Developing text encoders specifically tailored for Bengali could enhance the capabilities of language models in processing and generating Bengali text.
- **Bengali Knowledge Graph Dataset:** Building a comprehensive Bengali knowledge graph dataset across various domains would further strengthen the linguistic representation and capabilities of these models.
- **Future Work Focus:** Future research may involve specialized training of these language models on old Bengali literature, including works by Rabindranath Tagore and other renowned authors. The goal is to improve understanding and generation of archaic Bengali linguistic styles.



# Thank you!

