Massive Multilingual Knowledge Graph of Semantic Domains

Rajarshi Biswas, Arpan Datta, Nikhil Katiki, Rajib Kumar Mahato, Nitesh Wagh, Matthew A. Lanham
Purdue University, Krannert School of Management
biswas36@purdue.edu; datta41@purdue.edu; rmahato@purdue.edu; katiki@purdue.edu; nwagh@purdue.edu; lanhamm@purdue.edu

ABSTRACT

Languages are essential to human life. There are currently around 7151 languages being spoken. According to UNESCO Atlas of the World's Languages in Danger, between 1950 and 2010, 230 languages went extinct and as of 2018, a third of the world's languages have fewer than 1,000 speakers left.

We are creating a multilingual database comprising of 500+ languages and mapping them to semantic domains. A graph database of words in different languages is mapped to their respective semantic domains creating a diverse list of words and phrases that will enable simpler literature translation. This will enable-

- Rapid development of linguistic resource (like dictionaries) in local languages.
- Development of new translation and NLP techniques that rely on word lists and topic modeling techniques.

INTRODUCTION

English words are mapped to their corresponding semantic domains and subsequently map words from other languages with the same/similar meaning to the respective semantic domain. This gives us a cluster of word lists and groups of words which is useful in creating rapid dictionaries and variety of downstream NLP tasks. To achieve such related mapped structure of words and their languages, regions, and semantic domains. DGraph database is being used. It extensively maps and segregates relations between each data point.

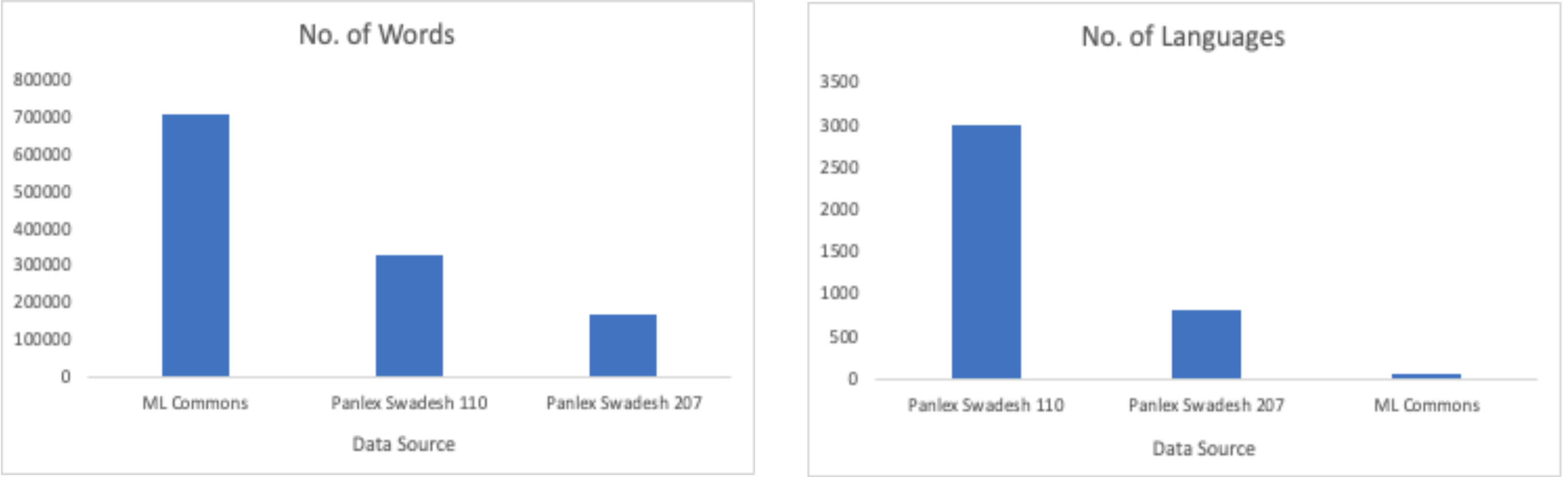

Fig 1. Data Composition

RESEARCH QUESTIONS

- Can existing multilingual dictionaries be combined with semantic domains to create a massively multilingual dictionary arranged in semantic domains?
- Can the massively multilingual dictionary transformed into a DGraph database?
- Can the data in the graph database be used to detect/predict the semantic domains for unmapped words/phrases?

METHODOLOGY

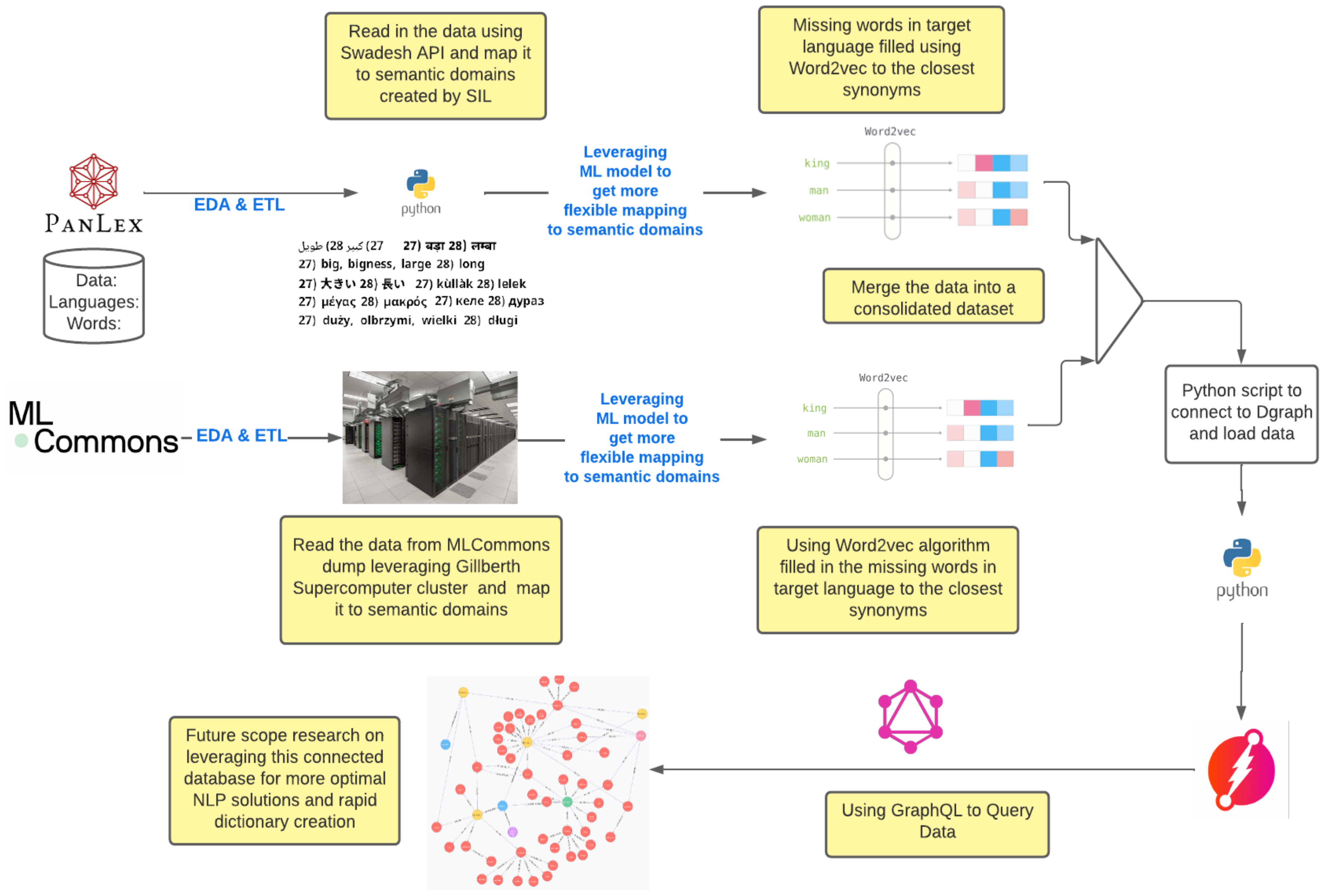
Read in the data using Swadesh API and map it to semantic domains created by SIL
Missing words in target language filled using Word2vec to the closest synonyms
PanLex
Data:
Languages:
Words:
EDA & ETL
python
Leveraging ML model to get more flexible mapping to semantic domains
Word2vec
king
man
woman
27) كبير 28) طويل 27) बड़ा 28) लम्बा
27) big, bigness, large 28) long
27) 大きい 28) 長い 27) kùllàk 28) lelek
27) μέγας 28) μακρός 27) кеłe 28) дураз
27) duży, olbrzymi, wielki 28) długi
Merge the data into a consolidated dataset
Python script to connect to Dgraph and load data
ML Commons
EDA & ETL
Leveraging ML model to get more flexible mapping to semantic domains
Word2vec
king
man
woman
Read the data from MLCommons dump leveraging Gillberth Supercomputer cluster and map it to semantic domains
Using Word2vec algorithm filled in the missing words in target language to the closest synonyms
python
Future scope research on leveraging this connected database for more optimal NLP solutions and rapid dictionary creation
Using GraphQL to Query Data

Fig 2. Methodology

STATISTICAL RESULTS

800K
Words

3K
Languages

9K
Word2Vec

4K
SemDom

- We are creating a database consisting 800K words mapped to 4K semantic domains
- The languages covered are 3000
- We labelled 25% of the unlabeled knowledge base of words using ML techniques

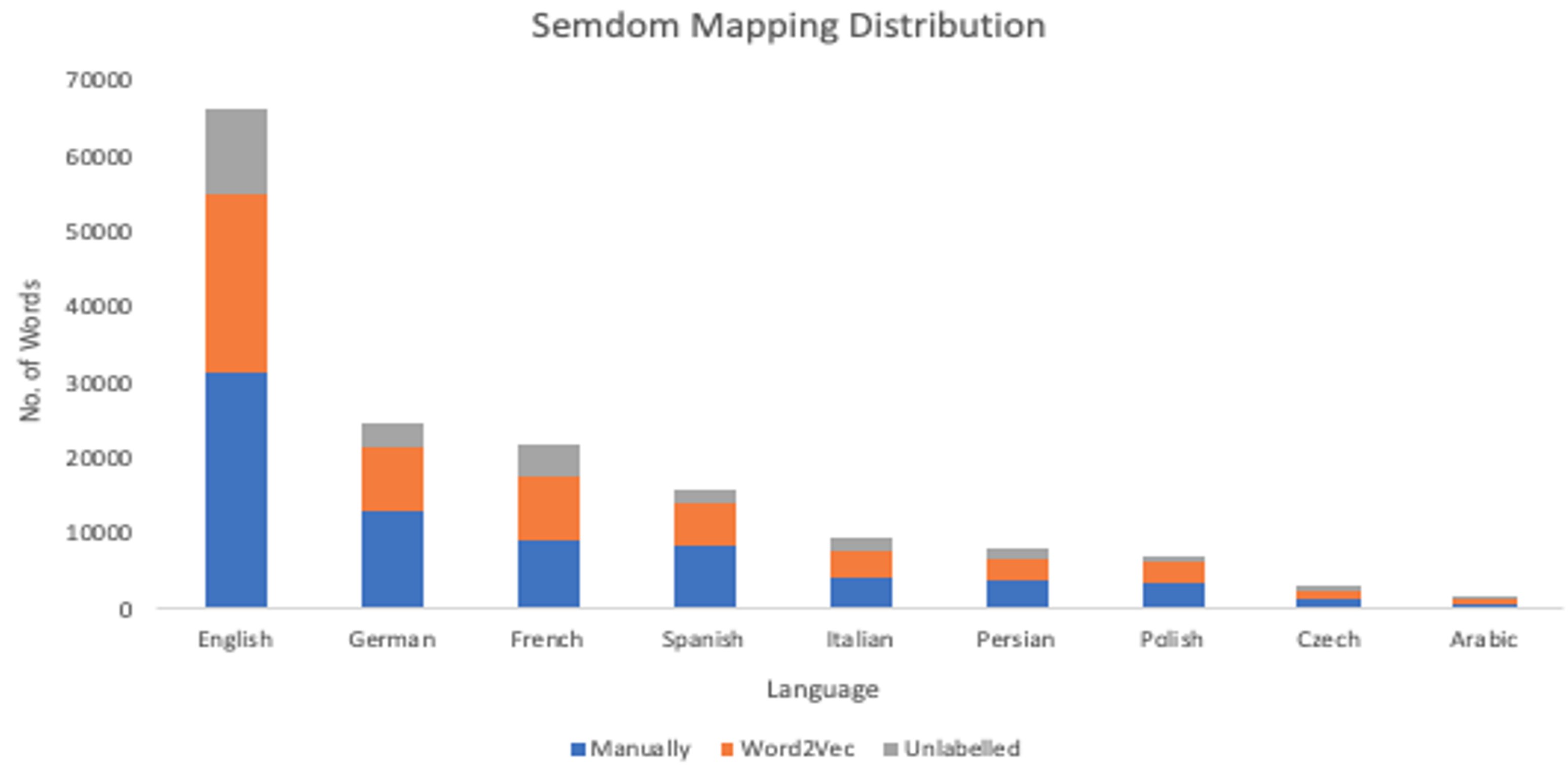

Fig 3. Top 9 Languages Direct translation vs Word2vec

Expected Business Impact

- Create a massively multilingual repository of words arranged in semantic domains for use in dictionary creation tasks and other downstream NLP tasks (for e.g., Dialogue Systems)
- Develop a foundational database for NLP practitioners to innovate using advance ML and DL techniques
- Enabling prediction of key terms in low resource, minority languages using words that are related (linguistically) in the graph

Business Case: Understanding the variations of word *Sun*.

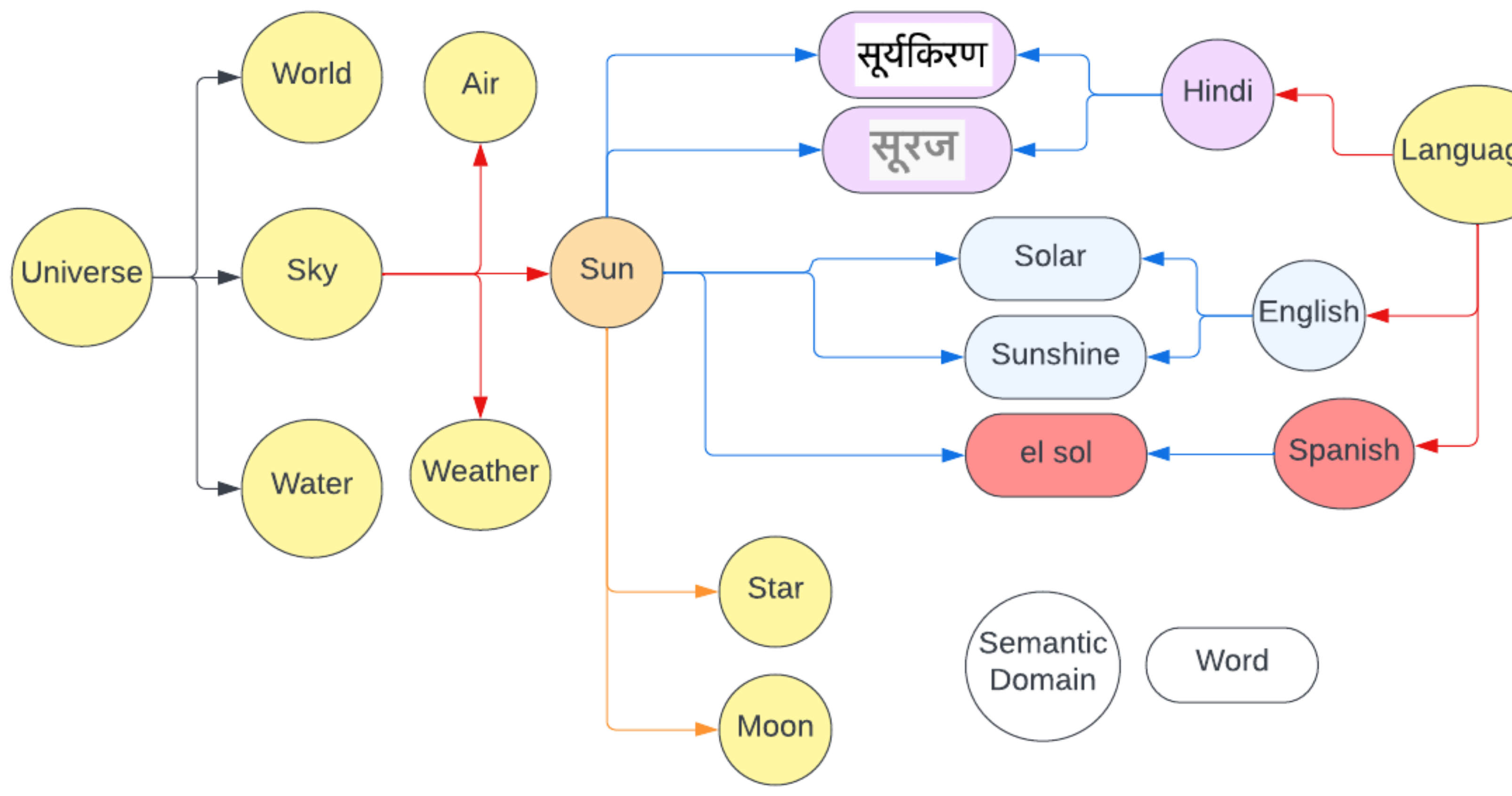

Fig 4. Representation of Graph Database

CONCLUSIONS

- We have successfully created a consolidated multilingual dictionary mapped to semantic domains leveraging machine learning techniques and existing knowledge base of words in different languages
- We built a process to load the consolidated data mapped to semantic domains to DGraph which has the scope of adding more words and languages in the future as per future requirements
- We were able to query the similar words in different languages from same semantic domains in the DGraph, this will enable to detect/predict the semantic domains of unmapped words

ACKNOWLEDGEMENTS

We would like to thank our professors Prof. Matthew Lanham and Yang Wang. We also would like to appreciate the guidance and mentorship from our industry specialist Daniel Whitenack for this project providing us with key suggestions and insights.