NCCS Research Journal, Volume 3, No. 1 October 2024, ISSN NO: 2822-1605

Received Date: June 2024 Revised: August 2024 Accepted: October 2024

Information Extraction from a Large Knowledge Graph in the Nepali Language

Dadhi Ram Ghimire

Assistant Professor

Department of Statistics and Computer Science

Patan Multiple Campus

dadhi.ghimire@pmc.tu.edu.np

Sanjeev Panday, PhD

Associate Professor

IOE, Pulchowk,

Email: sanjeeb@ioe.edu.np

Aman Shakya, PhD

Assistant Professor

IOE, Pulchowk,

Email: aman.shakya@ioe.edu.np)

Abstract

Information is abundant on the web. The knowledge graph is used for organizing information in a structured format that can be retrieved using specialized queries. There are many Knowledge graphs but they differ in their ontologies and taxonomies as well as property types that bind the relation between the entities, which creates problems while extracting the knowledge from them. There is an issue in multilingual support. While most of them claim to be multilingual they are more suitable for querying in the English language. Most of the existing knowledge graphs in existence are based on Wikipedia Infobox. In this work, we have devised an information extraction pipeline for retrieving knowledge in Nepali Language from Wikidata using SPARQL endpoint. Queries based on Wikipedia infobox has more accurate responses than the Queries based on the paragraph content of Wikipedia articles. The main reason behind that is that the information inside the paragraph is not linked properly in the Wikipedia infobox.

Keywords: Question Answering, Semantic Network, Knowledge Graph, WikiData, SPARQL

Information Extraction from a Large Knowledge Graph in the Nepali Language

Knowledge graphs, a structured organization of knowledge, have garnered significant attention in recent times from both academic and industry research departments. An organised representation of facts made up of entities, relationships, and semantic descriptions is called a knowledge graph. Relationships show the link between things, and entities themselves can be both concrete objects and abstract ideas. Semantic representations of entities and their relationships include types and characteristics that have clear definitions. Often used graphs with nodes and relations having qualities or attributes are called property graphs or attributed graphs. Except for a little distinction, knowledge graph and knowledge base are interchangeable terms. Given its graph structure, a knowledge graph can be thought of as a graph. When it comes to formal semantics, it can be viewed as a foundation of knowledge enabling fact-based interpretation and deduction (Ji et al., 2022). Structured facts are used in knowledge graphs (KGs) to explain the real world. A fact can be defined as a SPO triple (Subject, Predicate, Object) made up of two things and the connection that connects them. An example of a fact would be (खोला, बस्छ, माछा).

Knowledge graphs are an effective way to show links between entities for organising data in structured format. Natural language processing, recommendation systems, semantic search, question answering and other fields have uses for them. But there are a number of obstacles that must be overcome in order for knowledge graphs to effectively represent information, which affects their usefulness and effectiveness. There are many Knowledge graphs but they differ in their in their ontologies and taxonomies as well as properties types that binds the relation between the entities, which creates problem while extracting the knowledge from them. There is issue in multilingual support. While most of them claim to be multilingual they are more suitable for querying in English language. This research focuses on queries against WikiData for extracting information in Nepali Language.

Background

There are many structured knowledge graphs which are freely accessible in the web. The study of these graphs can provide an insight of knowledge representation in

large knowledge graphs, which will be helpful in creating a structured knowledge graphs in Nepali Language.

Ontology and Taxonomy

A set of concepts inside a domain and the connections among them are formalized in an **Ontology**. To simulate the knowledge inside that field, it offers a common vocabulary. A domain's entity types, as well as the attributes and connections between them, are specified using ontologies. Property, constraint, and class information are included here. (McHugh, 2023)

Components of Ontology:

Classes: The main things in the domain (e.g., Person, Book) are called classe or concepts.

Qualities: Qualities or characteristics of the classes (e.g., population, age, title).

Relations: The connections between classes and properties (e.g., a city is situated in a country, an author writes a book). Examples of the classes in particular are called instances (for example, "John Doe" is an instance of the class Person).

A **taxonomy** is a kind of hierarchical classification that arranges ideas according to parent-child relationships and frequently takes the form of a tree structure. To make it easier to navigate, comprehend, and retrieve information, taxonomies are used to group and categorise items. (McHugh, 2023)

Components of Taxonomy:

Nodes: The distinct components or groups that make up the taxonomy.

Edges: The nodes' connections with one another, which show hierarchical relationships (for example, a book is a type of book, while a mammal is an animal).

Hierarchy Levels: Various abstraction levels, with higher levels denoting broader categories and lower levels more narrowly defined ones.

Frameworks for Knowledge Representation

Web Ontology Language (OWL)

Ontologies is created and shared on the web using the formal language OWL. Its goal is to make it possible to define and exchange web-based knowledge in an

organised, compatible way. It supports a wide range of operators to specify attributes, classes, and the connections between them, making it possible to express intricate relationships between concepts. It is developed on top of RDF and RDFS. In addition to that, OWL adds more vocabulary and semantic powers to these frameworks, enhancing their capacity to share data and work together across many systems and domains. OWL facilitates reasoning regarding the connections between concepts. Applications are thus able to deduce logical inferences from the data through automated inference of new knowledge based on the defined ontology (OWL 2 Web Ontology Language Document Overview (Second Edition), n.d.).

Resource Description Framework

Resource Description Framework (RDF) is a graph-based data model proposed for the realisation of the Semantic Web vision and key format of the Linked Data publication strategy. It makes use of triples, or sentences of the format subject – predicate – object, in which the subject is an entity (a product, a company, etc.), the predicate is a characteristic of the entity (a product's price, the location of a company, etc.), and the object is the predicate's value for the particular subject (a product, a company, etc.). Triples are utilised to relate anonymous resources (blank nodes) or uniform resource identifiers (URIs) to other URIs, blank nodes, or constants (Literals) (Papadaki et al., 2023).

A collection of classes with specific characteristics that are based on the RDF extensible knowledge representation data model make up the RDF Schema, a unique vocabulary. Although RDF use URIs to uniquely identify resources, it is not semantically expressive, hence its goal is to arrange RDF resources. It makes use of properties to establish relationships between entities in a class and to represent constraints, and classes to show where a resource belongs (Papadaki et al., 2023).

Schema.org

Schema.org provides a common vocabulary for annotating data on web pages, which is essential to the creation of Knowledge Graphs. It creates a vocabulary that websites may use to define their content, such as Person, Event, and Location. In addition to helping websites embed rich information within their code, it enables search engines and other applications to comprehend the meaning and relationships between various pieces of information, identify entities, and enable the creation of connections

between related entities across various websites, all of which contribute to the overall knowledge graph (Iliadis et al., 2023) .

Existing Knowledge Graphs

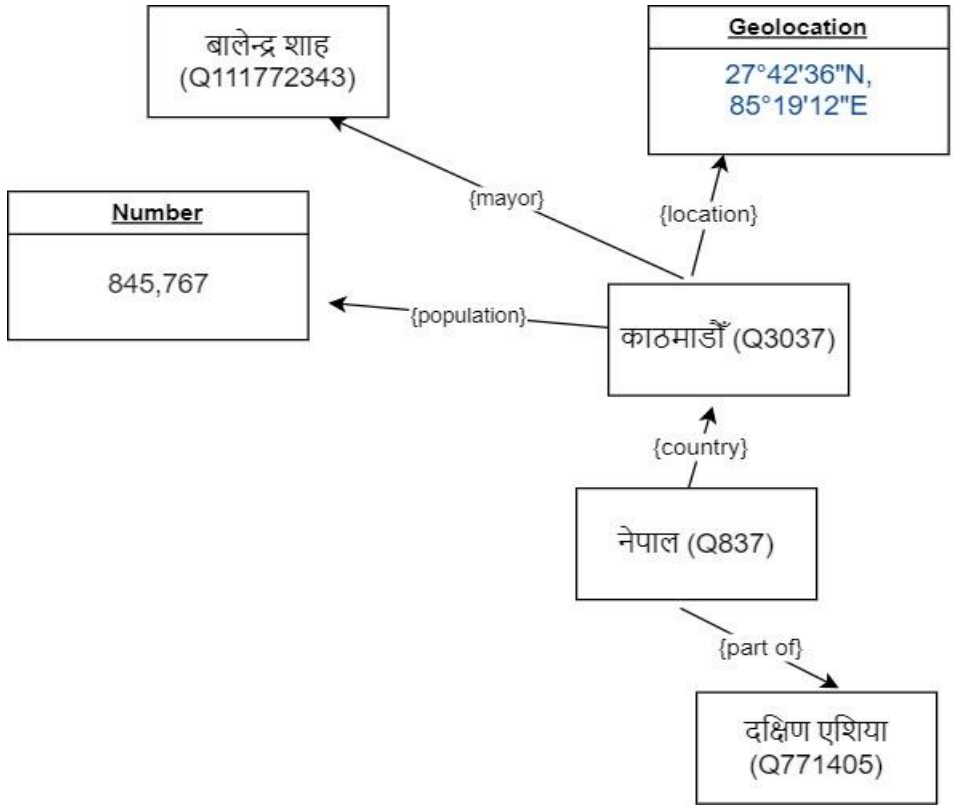
Wikidata

One of the largest general-purpose KBs nowadays is Wikidata. It offers a plethora of information on almost every topic spoken in everyday conversation, with over 100 million entities and 1.4 billion data about them. Since each entity has an abstract identifier (like Q83396), the identifiers are time- and language-invariant. The project has tens of thousands of contributors.

One of the largest general-purpose KBs nowadays is Wikidata. It offers a plethora of information on almost every topic spoken in everyday conversation, with over 100 million entities and 1.4 billion data about them. Since each entity has an abstract identifier (like Q83396), the identifiers are time- and language-invariant. The project has tens of thousands of contributors. (Suchanek, 2024)

Figure 1

Interconnected entities along with their properties and values in Wikidata



The primary components of the Wikidata repository are items, each of which has a label, a description, and an infinite number of aliases. A Q and a number serve as a unique identifier for an item; Douglas Adams is one example (Q42). Statements are made up of a property and a value that explain certain attributes of an item. In Wikidata, properties are denoted by a P and a number, for example, educated at (P69) (*Wikidata:Introduction - Wikidata*, 2024).

Babel Net

BabelNet is a multilingual semantic network Converging diverse resources including WordNet, Wikipedia, Wikidata, Wiktionary, and numerous others. Completing the image of the lexical and semantic knowledge gleaned from the integrated resources is achieved by integrating disparate pieces of information, much like in a jigsaw puzzle.

The concept of a synset, or the collection of synonymous words or senses that can be used to describe the same meaning in a particular language, is the foundation upon which BabelNet bases its representation of each meaning. For instance, in WordNet, the named entity NEW YORK is defined as the set {New York, New York City, Greater New York}, while the notion of DOG is represented by the set of terms {dog, domestic dog, Canis familiaris}. This idea is expanded by BabelNet to encompass equivalent lexicalizations across several languages (Navigli et al., 2021).

YAGO

YAGO is a comprehensive knowledge base that emphasises entities, facts, and their connections while presenting information in an organised manner. It combines a great deal of data from several sources, such as WordNet, Wikipedia, and GeoNames. Numerous subjects are covered, such as geography, history, biology, and culture. In order to provide a formal framework for expressing entities and their properties, YAGO organises knowledge into a hierarchical ontology. YAGO combines data from several languages, enabling the representation and reasoning of cross-lingual knowledge. YAGO uses automated techniques to extract information from unstructured text sources, which makes it possible for the knowledge base to grow and update automatically.

YAGO enhances our comprehension of the connections between distinct

concepts by capturing a wide range of semantic links between entities, including is-a, part-of, located-in, and many more (Tanon et al., 2020).

DBpedia

DBpedia is a knowledge graph extracted from Wikipedia article's infoboxes. It enables users to connect the data to other online databases, run sophisticated queries on it, and use it for a variety of purposes. The infoboxes hold organised information regarding the topic of the article, including a person's birthdate, a city's population, or a product's characteristics. DBpedia is well known member of the Semantic Web and Linked Data initiatives, which seek to build an easily machine-processable web of interconnected data. The infobox templates are parsed throughout the extraction process, and the data is then transformed into RDF triples (About DBPedia - DBpedia Association, 2021).

ConceptNet

A labelled, weighted edge (an assertion) links words and phrases of natural language (terms) in ConceptNet. ConceptNet additionally depicts connections among information sources. In order to describe a relationship regardless of the language or source of the terms it connects, ConceptNet employs a closed class of chosen relations like IsA, UsedFor, and CapableOf.

` The most significant source of input for ConceptNet is Wiktionary, which contributes 18.1 million edges and is primarily responsible for its extensive multilingual vocabulary. ConceptNet comprises over 21 million edges and over 8 million nodes. Its English vocabulary comprises about 1,500,000 nodes, and it contains at least 10,000 nodes in 83 languages (Speer et al., 2017).

The CommonSense Knowledge Graph

CSKG is an extensive knowledge graph that includes common sense information. Information from multiple sources, including ConceptNet, ATOMIC, and others, is integrated to create the graph. With a focus on a broad range of subjects and situations, CSKG seeks to offer an organised representation of common-sense knowledge akin to that of humans. The article addresses the development process, assessment techniques, and possible uses of CSKG in a number of fields, including artificial intelligence, machine learning, and natural language understanding. It includes a wide range of topics, such as ideas, methods, and even the relationships between

words. It is set up in the form of a hyper-relational graph, which is a sophisticated method of displaying information linkages (Ilievski et al., 2021).

ATOMIC

Commonsense information about human interactions and behaviour is captured in the ATOMIC knowledge graph. It portrays an organised set of atomic-level occurrences, each with a subject, verb, and object as well as extra contextual details. The abbreviation ATOMIC is "The Atlas of Technologically Mediated Conversations." By offering a wealth of CommonsSense knowledge, it was designed to support research in artificial intelligence, machine learning, and natural language understanding. The graph includes more than 300,000 distinct textual descriptions of events that span a variety of situations and human behaviours. To facilitate more in-depth analysis and interpretation, each event is annotated with extra details like temporal information, intents, and effects (Sap et al., 2019b).

Related Works

The collaboratively editable repository QAWiki, which compiles questions in several natural languages along with the associated structured enquiries, serves as the foundation for Templet. Templet creates templates from question-query pairs on QAWiki. The user can generate a concrete question, query, and results by typing a question in natural language, choosing a template, and then using autocompletion to choose the entities they want to enter into the template's placeholders (Suárez & Hogan, 2023).

The issue of open domain factoid-based question responding, in which the response takes the form of a single word or brief phrase, is addressed in this paper. It offers a productive approach to question analysis and knowledge base-based suitable answer specification. Before anything else, a thorough linguistic study of a user-specified question is carried out. The question's primary triplets are developed, and its dependencies are indicated. To obtain the correct response to the inquiry, the system formulates a SPARQL query using the Wikidata inquiry Service API (Ploumis et al., n.d.).

Methodology

The structured information stored in the knowledge graph can be extracted in various ways. The most common approach is to use the SPARQL Queries against the knowledge base. The source of information of most of the large knowledge graphs is the Wikipedia infobox. The figure 2 below presents a simple pipeline to extract knowledge in Nepali Language.

SPARQL

Data stored in the Resource Description Framework (RDF) format can be queried and altered using the robust SPARQL query language and protocol. The acronym SPARQL represents SPARQL Protocol and RDF Query Language. **Triples** (subject-predicate-object statements) make up RDF data, which is specifically intended for querying. Triple pattern matching against RDF data is the fundamental operation of SPARQL queries (SPARQL 1.1 Query Language, n.d.).

Many kinds of queries are supported by SPARQL:

SELECT: Takes raw information out of the RDF database. Constructs new RDF graphs by using the query results as a basis.

ASK: Provides a Boolean answer indicating if the pattern of the question matches or not.

DESCRIBE: Provides an RDF graph that details the resources that were located.

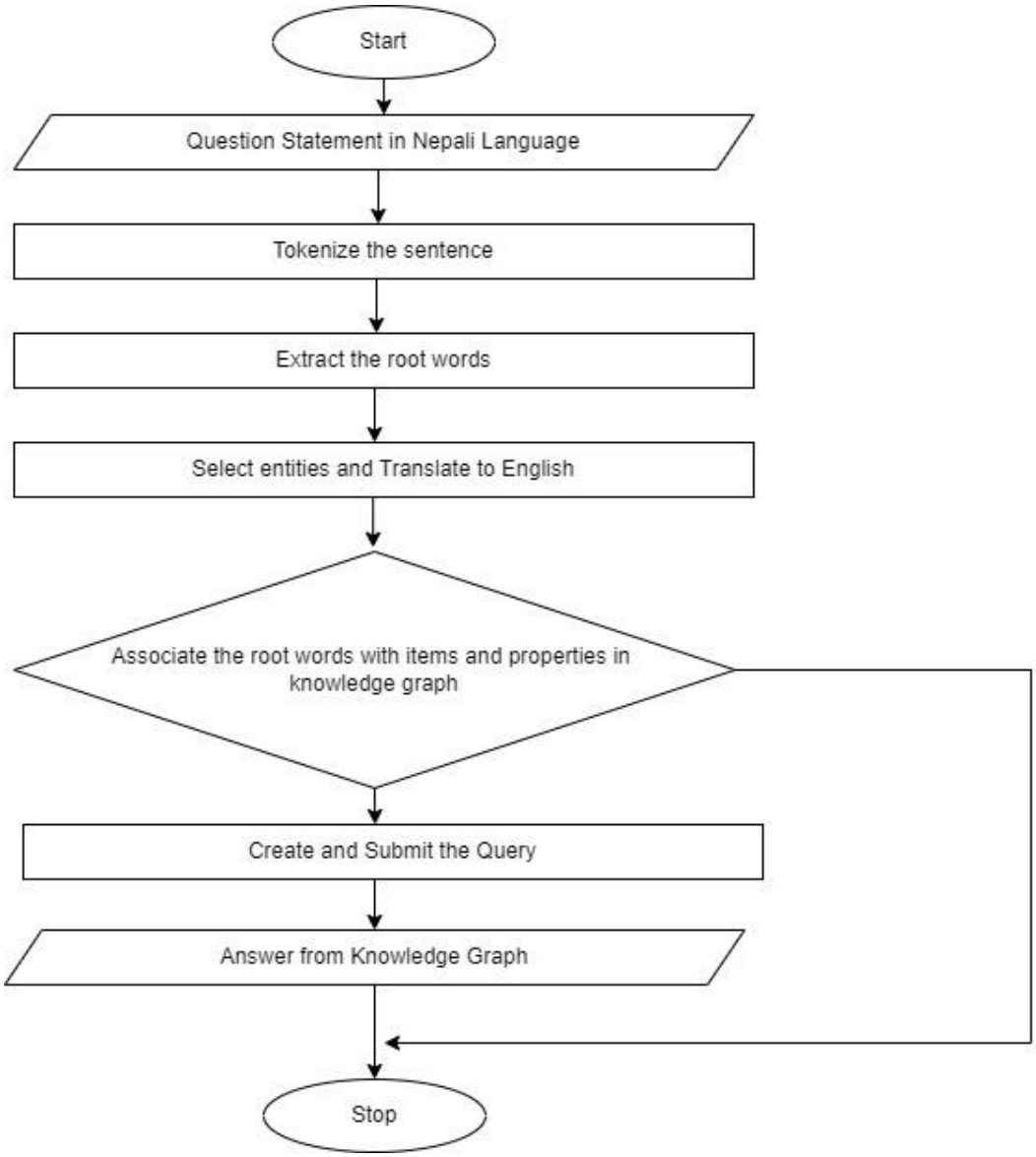
Example:

```
SELECT ?name WHERE
{
  ?person rdf:type foaf:Person .
  ?person foaf:name ?name .
}
```

This query searches for triples where the subject (**?person**) is of type **foaf:Person** and has a **foaf:name** property, returning the names.

Figure 2

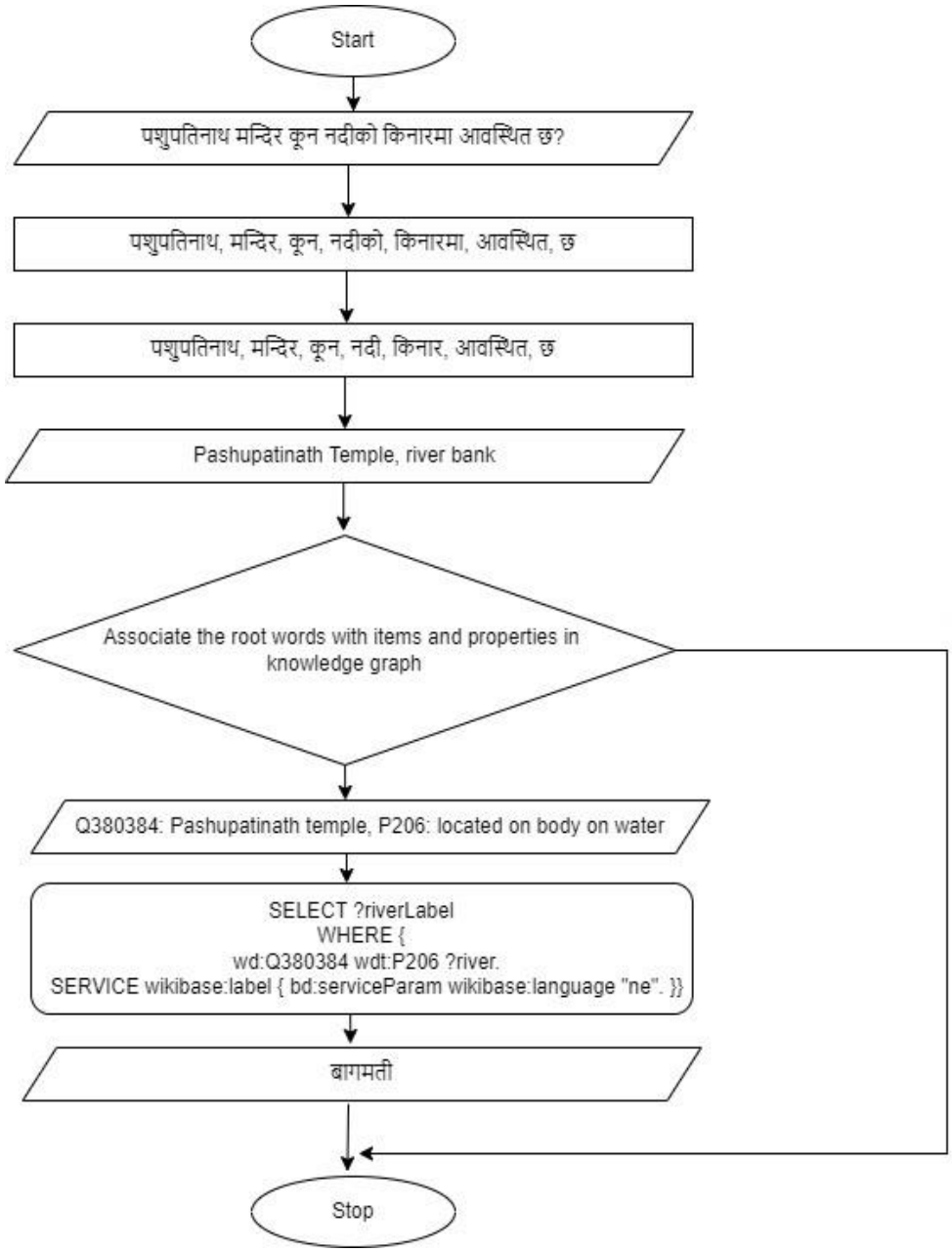
Query Extraction Pipeline



Knowledge Extraction from Wikidata against Query in Nepali Language is visualized in figure 3 using a working example.

Figure 3

Query Extraction Pipeline in Wikidata example



The question in Nepali language is tokenized first and the root words are extracted by removing inflections and postpositions. Then the subject and predicates are selected (manually done in this study) and these are matched against the Wikidata item ID and Wikidata properties then SPAQRL query is created and executed and results are noted down.

Implementation tools

Implementation is done using Python using the libraries such as SQLWrapper, Pandas etc in Google Colab. Wikidata SPARQL Queries were generated using RDF Framework.

Data Set

Two types of data set were prepared, 30 questions were generated manually by exploring infobox of different Wikipedia articles in Nepali Language Such as नेपाल (Q837), पशुपतिनाथ मन्दिर (Q380384), बागमती नदी (Q4461769) etc.

पशुपतिनाथ मन्दिर कून नदीको किनार मा आवस्थित छ ?

पृथ्वीनारायण शाहको जन्म कैले भएको थियो ?

Similarly, 30 Questions were generated manually by manually exploring paragraph content of Wikipedia Articles in Nepali Language.

आधुनिक नेपालको राष्ट्रनिर्माताको रूपमा कसलाई चिनिन्छ?

नेपालमा प्रजातन्त्रको लागि क्रान्ति कहिले भएको थियो?

Implementation Details

Tokenizing the input question sentence

```
from nepalitokenizers import WordPiece
text = "पृथ्वीनारायण शाहको जन्म कैले भएको थियो?"
tokenizer_wp = WordPiece()
tokens = tokenizer_wp.encode(text)
print(tokens.ids)
print(tokens.tokens)
print(tokenizer_wp.decode(tokens.ids))
```

'पृथ्वीनारायण', 'शाहको', 'जन्म', 'कैले', 'भएको', 'थियो'

Generating the root words

['पृथ्वीनारायण', 'शाह', 'जन्म', 'कै', 'भए', 'थियो', '?']

Purging and Translating the words (Manual)

Prithvi Narayan Shah, Birth

Matching the item ID and Property in Wikipedia

wd:Q574450 wdt:P569

Creating the SPARQL Query

```
SELECT ?dob WHERE
{
wd:Q574450 wdt:P569 ?dob.
}
```

Result

7 January 1723

Result Analysis

Table 1 shows the information regarding responses from Wikidata Knowledge graph in Nepali and English language.

Table 1

Response to queries

Questions Type	Total Questions	Answer in Nepali Language	Answer in English Language	No Answer
Questions from Wikipedia Infobox	30	23	3	4
Questions from Wikidata Paragraph Content	30	4	8	18

Similarly, table 2 shows the information regarding correct and incorrect answer to the questions presented to the Wikidata Knowledge Graph.

Table 2

Information on Correct and Incorrect Answer

Questions Type	Total Questions	Correct Answer	Incorrect Answer	No Answer
Questions from Wikipedia Infobox	30	25	3	2
Questions from Wikidata Paragraph Content	30	10	2	18

From above two table we can observe that, the responses to Queries based on Wikipedia infobox has more accurate responses than the Queries based on the paragraph content of Wikipedia articles. The correctness of answer was checked manually against the fact present in Wikidata infobox. The main reason behind that is that the information inside the paragraph is not linked properly in the Wikipedia infobox. For an instance while querying for the answer of "नेपालको राष्ट्रिय फुल कुन हो"?

, the entities ''नेपाल'' and ''फुल'' have respective articles in Wikipedia however these are not linked properly through the infobox. In table 2 we can see that some of the responses of the query in Nepali language is returned in English language as the number of articles in Wikipedia about different entities in Nepali Language is far below compared to articles in English Language.

Conclusion

This study gave valuable insight into the structure and source of knowledge of large KGs. Infobox of Wikipedia articles is the major source of knowledge graphs such as Wikidata, DBpedia, YAGO, Babel Net etc. Some KGs offer different ontologies and Taxonomies and link to other KGs. The SPARQL query based on the information listed in the Wikipedia infobox seems to be more effective than the question based on information in the paragraph content of Wikipedia articles. Similarly, the amount of knowledge in this knowledge graph in the Nepali Language is minimum compared to other language.

References

About DBPedia - DBpedia Association. (2021, March 5). DBpedia Association. https://www.dbpedia.org/about/

Iliadis, A., Acker, A., Stevens, W. M., & Kavakli, S. B. (2023). One schema to rule them all: How Schema.org models the world of search. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24744

Ilievski, F., Szekely, P., & Zhang, B. (2021). CSKG: The CommonSense Knowledge Graph. In *Lecture notes in computer science* (pp. 680–696). https://doi.org/10.1007/978-3-030-77385-4_41

Ji, S., Pan, S., Wang, Z., Marttinen, P., & Yu, P. S. (2022). A survey on Knowledge Graphs: Representation, acquisition, and Applications. *IEEE Transactions on*

Neural Networks and Learning Systems, *33*(2), 494–514. https://doi.org/10.1109/tnnls.2021.3070843

McHugh, J. (2023, August 17). *Taxonomies, ontologies, semantic models & knowledge graphs*. BigBear.ai. https://bigbear.ai/blog/taxonomies-ontologies-semantic-models- knowledge-graphs/

Navigli, R., Bevilacqua, M., Conia, S., Montagnini, D., & Cecconi, F. (2021). Ten years of BabelNet: A survey. In *IJCAI* (pp. 4559-4567). International Joint Conferences on Artificial Intelligence Organization.

Papadaki, M., Tzitzikas, Y., & Mountantonakis, M. (2023). A Brief Survey of Methods for Analytics over RDF Knowledge Graphs. *Analytics*, *2*(1), 55–74. https://doi.org/10.3390/analytics2010004

Ploumis, T., Perikos, I., Grivokostopoulou, F., & Hatzilygeroudis, I. (n.d.). *A Factoid based Question Answering System based on Dependency Analysis and Wikidata* (Vol. 2, pp. 1–7). https://doi.org/10.1109/iisa52424.2021.9555551

Sap, M., Bras, R. L., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., & Choi, Y. (2019). ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. *Proceedings of the . . . AAAI Conference on Artificial Intelligence*, *33*(01), 3027–3035. https://doi.org/10.1609/aaai.v33i01.33013027

SPARQL 1.1 Query language. (n.d.). Retrieved May 25, 2024, from https://www.w3.org/TR/sparql11-query/

Speer, R. E., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. *Proceedings of the . . . AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.11164

Suárez, F., & Hogan, A. (2023). Templet: A Collaborative System for Knowledge Graph Question Answering over Wikidata. *WWW '23 Companion: Companion Proceedings of the ACM Web Conference 2023*, 152–155. https://doi.org/10.1145/3543873.3587335

Suchanek, F. (2024). YAGO 4.5: A Large and Clean Knowledge Base with a Rich Taxonomy. *In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Tanon, T. P., Weikum, G., & Suchanek, F. M. (2020). YAGO 4: A reason-able knowledge base. In *Lecture notes in computer science* (pp. 583–596). https://doi.org/10.1007/978-3-030-49461-2_34

Wikidata:Introduction - Wikidata. (2024, January 24). Retrieved May 7, 2024, from https://www.wikidata.org/wiki/Wikidata:Introduction