A Semi- Automatic Framework for the Development and Analysis of Selective Natural Language Ontologies

Muhammad Ishaq [0000-0003-1963-5041] , Asfandyar Khan [0000-0001-5174-0736]
Arshad Khan.
The University of Agriculture, Peshawar, Pakistan
drmishaq@aup.edu.pk

Abstract
The goal of the next generation World Wide Web is machine readability through linked databases. To improve web search, integration, and mining in local languages like Urdu, there is a growing need to develop ontologies and vocabulary in these languages. The majority of people use the web in local languages for agriculture, social media interaction, news, etc. How to create agents for the integration of web data. In our country, the literacy ratio is very low and IT literacy is negligible. More comprehensive information for its target audience is only possible through the World Wide Web. Our first target is to improve and enhance the use of social media and the web in local languages. That will encourage its constructive use in Urdu for society and the economy.
The Web in natural languages is the source of income for small and medium enterprises. The semantic web is concerned with linked databases and structured data. In this work, we are focused on some selected ontologies to be translated into natural languages. Expertise in Ontology Engineering helps us in job production. Ontology Engineering has extensive freelancing opportunities. Only if the web is correctly interpreted in regional languages is an economic boost achievable. A standardized foundation for data sharing and reuse on the internet is provided by the Semantic Web. In other words, a group of standards and technology that enables computers to comprehend the semantics (meaning) of material on the Web.

Keywords: Natural Language Ontologies, Ontology Engineering, Ontology Development, Semantic Web.

1. Introduction

Semantics is a logic concerned with meaning or how to make possible the web of meaning. Syntax means how we say something, and semantics is the real meaning behind the said words.
Properly linked nodes in a triple on the web, as we draw the entity relationship diagram of a relational DBMS. Structured data means data with proper predefined RDF and OWL based data models.
To improve web search, integration, and mining in local languages like Urdu, there is a growing need to develop ontologies and vocabulary in these languages. A majority of people use the web in their local languages for social media, news, food and literature. In this case, natural language processing (NLP) works as an intermediate layer between natural languages and ontologies.
A robust framework for domain-specific Unified natural language vocabularies (ontologies) can ease human effort in searching, mining, and integration. Semantic is only possible through the linking of data sets. Linking means connecting the data sets of unstructured web data to a certain standard with suitable tools and technologies. Common standards are the RDF, OWL, and the SPARQL query language. Linked

 doi:10.20944/preprints202211.0201.v1

databases are only possible through the development of ontologies or vocabulary for any specific domain. The goal is to integrate and interlink web data. It includes data from different websites. The semantic web allows us to combine a term's database and vocabulary (ontologies) into a single point with a unique URI. In any semantic system, each individual term has its own link called an IRI. The Layered Structure of Semantic Web is shown in the Figure 1.

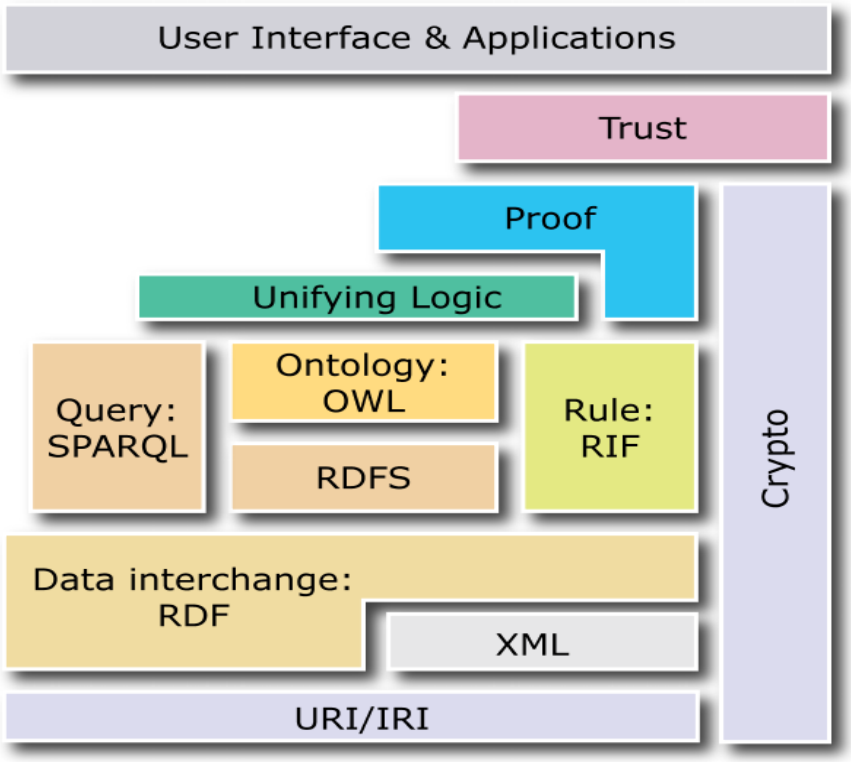

Fig. 1. The Layered Structure of the Web of Meaning.

1.1 RDF and OWL

RDF enables you to connect materials (concepts), therefore you could say that (Karthik is a person). Consider it to be a directed graph. However, since things cannot be classified, it is impossible to claim, for instance, that a person belongs to the subclass of human beings, etc.

More constraints can be added to your knowledge representation using OWL. It allows you to set limits to your properties and divides properties (relationships) into object and data properties.

The development of a Web-based online ontology development framework using Protégé with a SPARQL query retrieval system is the net outcome of this work. An overwhelming number of web applications and web-based software systems dominate the software industry. There is a boom in Semantic Sciences and Ontology Engineering jobs. Almost all big corporate, IT, and research firms are busy or tend to develop

 doi:10.20944/preprints202211.0201.v1

ontologies in relevant domains. To facilitate our use, machine readable is the only solution. A data model is like the traditional relational model. Therefore, there has to be a way (a model) to represent knowledge on the Web and it should be accepted by all websites on the WWW. There is a need to have a uniform way of creating such statements. The web contents or statements should not be arbitrary and there should be a standard mechanism to do it. The contents should be linked with proper relationships among different relevant terms. OWL and RDF are common frameworks for all data models in linked databases. Each data model supports the automatic translation of logical database diagrams into physical database designs.

For readability, the machines need the web to be defined in linked database format. To build a data integration agent, we need to make some changes to the websites. Each statement collected by our agent represents a piece of knowledge. The goal of the next generation World Wide Web is machine readability through linked databases. Semantics is a logic concerned with meaning or how to make possible the web of meaning. Syntax means how we say something, and semantics is the real meaning behind the said words. Give proper meaning to information and make it available for proper machine reasoning and how to draw meaningful inference. The semantic web is concerned with linked databases and structured data. Properly linked nodes in a triple on the web, as we draw the entity relationship diagram of a relational DBMS. Structured data means data with a predefined data model. The semantic web is a new paradigm of knowledge management.

We need the WWW for searching, data integration, and mining the web. To facilitate our use, machine readable is the only solution. We direly need a linguistic expert for proper and correct NLO development. We are engaged in the translation of English ontologies. Each term should be translated according to the rules of Natural Language (NL).

In our case, the NL is Urdu. There are a lot of reasons for using the Protégé editor. The Editor is extensively used, and the project is run by Stanford.

2 Related Work

Agricultural and biological ontologies translated to natural languages improve relevant web-based informatics [1]. The farmer’s community gets up-to-date knowledge and expertise about crop protection and harvest.

Bioinformatics ontologies like Multiple Sequence Alignment Ontology (MAO) catch the interest of researchers in this interdisciplinary field of research [2].

The contents should be linked with proper relationships among different relevant terms. OWL and RDF are common frameworks for all data models in linked databases.

The word "property" is used to describe connections between people or even classes. RDB uses relationships to correlate different entity kinds or entities. Slots, roles, or traits of each idea describe its characteristics and attributes[3]. Facets (interfaces) or role limitations are constraints on slots. A form of relationship between individuals, such as having a parent, a pet, or a service number, is described as a relationship between individuals (and data). Data properties and object properties are the two categories.

In the area of conversation or the area of interest to us, people represent real-world goods or services. During the ontology creation process, Protégé makes use of Unique Name Assumption (UNA) for individual nomenclature [4]. The Unique Name

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2022 doi:10.20944/preprints202211.0201.v1

Assumption (UNA) is not used by the Open Web Ontology (OWL), and two names could genuinely correspond to the same person. For example, "King Abdullah", "The King" and "Abdullah" might all refer to the same individual. In OWL, the similarity and difference between individuals must be explicitly stated. A diagrammatic illustration of some individuals in some domains is as under. Individuals are usually represented as diamonds in any relevant diagram as shown in the below Figure 2.

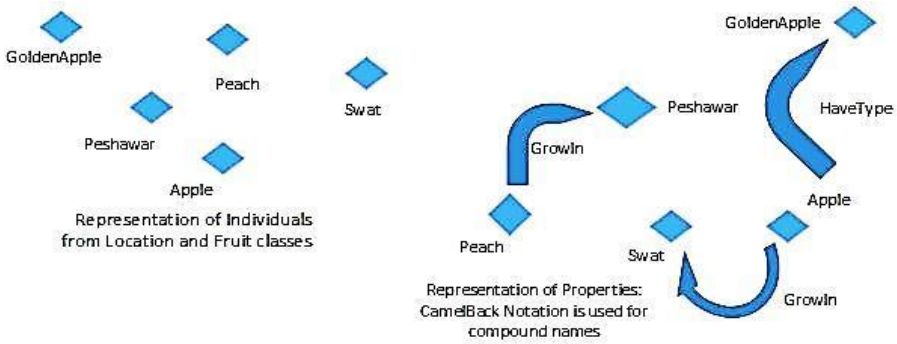

Fig. 2: Representation of Individuals and Triples showing Object Properties

A country is a type of thing. The country has many individuals using any number of protégé versions.

A subject, object, and predicate give rise to a triple. A subject or object may be any individual, class, or subclass. The predicate is actually the property [5].

If we develop a triple for all possible properties among individuals based upon some standards and rules, then we call it the ontology of that specific domain.

After proper definition of all triples and correct annotations, the ontology is uploaded to any unified ontology repository. After successful upload and merge, it becomes machine readable. Machine refers to web integration and a search agent.

Ontology means integrated and interlinked domain-specific data. It includes data from different websites. The semantic web enables us to unify the database and vocabulary (ontologies) of any term into a uniform point with the proper uniform resource identifier (URI). In linked databases, each term (ontology) and node possess a URI. A node can be a subject, object, or predicate. A URI means a unique identifier on web3, like the Uniform Resource Locator (URL) on web2.

3 Semi Automation of Natural Language Ontology Development

The majority of people in Pakistan have weak English language skills, and they are used to social media and news in Urdu. This work is an effort to improve the semantics of social media and news on the World Wide Web (WWW).

Semantics is a logic concerned with meaning or how to make possible the web of meaning. Syntax means how we say something, and semantics is the real meaning behind the said words.

 doi:10.20944/preprints202211.0201.v1

Properly linked nodes in a triple on the web, as we draw the entity relationship diagram of a relational DBMS. Structured data means data with proper predefined RDF and OWL based data models.

The development of a Web-based online ontology development framework using Protégé with a SPARQL query retrieval system is the net outcome of this work. The main objectives of the semantic web, like data interoperability and ontology reusability, can only be ensured through a domain-specific library or repository.

Any word or topic's ontology or vocabulary is truly incorporated as a linked database in the web3, also known as the semantic web. The usage of standardized formats, such as RDF and OWL, ensures syntactic interoperability, enabling applications to reuse data and link various types of data. All ontology editors make use of the same RDF and OWL. The key technology that permits and supports interoperability at the semantic level is represented by ontologies, which offer a formal conceptualization of the data that can be shared, reused, and aligned [5].

We have the knowledge and experience to work effectively in Protégé (developed by Stanford with NIH funding)[13]. It is quite good to develop ontology in Urdu.

.

3.1 **OBJECTIVES achieved**

1. Through investigation of relevant natural language ontologies.
2. Develop a standardized approach for the creation, editing, mapping and reusability of ontologies in local languages.
3. Social media and News are the specific domains for which we have to develop ontologies in Urdu (natural language).
4. To provide proof that searching, integration and mining of traditional web in local language is improved.
5. Web based interface for ontology development and query retrieval system will be deployed.

3.2 **Development PROCEDURE**

a) In the beginning we will develop ontologies in natural language (Urdu) for terminology used in social media using protégé. Standardization and comparison of our work with relevant ontologies on freely available and well reputed repositories like the Stanford ontology repository and other W3C recommended repositories.

The waste field of ontology engineering has a wide range of ontology editors[14]. There is a slight variation of Concepts and terminologies associated with each editor. In this work, we will stick to all standardized concepts, terminologies and techniques issued by the Stanford protégé work.

Ontology develop on local machine can be uploaded to any relevant repository. There are various techniques to judge the correctness and impact of our ontology. The improvements of semantics of relevant terms on World Wide Web are an indicator of the positive impact of any ontology. Protégé use XLST translator for the mentioned natural languages. To register and develop our own W3C recommended ontology repository or domain is our optimal goal for Urdu ontologies in the above mentioned application domain.

b) Majority of people in Pakistan have weak English language command, and they are used to social media and news in Urdu. This work is an effort to improve the semantics of social media and news on the World Wide Web (WWW).

 doi:10.20944/preprints202211.0201.v1

Only a domain-specific library or repository can guarantee the primary semantic web goals, such as data interoperability and ontology reusability [6]. Any word or topic's ontology or vocabulary is really included as a linked database in the web3 or semantic web [7]. The usage of standardized formats, such as RDF and OWL, ensures syntactic interoperability, enabling applications to reuse data and link various types of data. Every ontology editor makes use of the same RDF and OWL. The key technology that permits and supports interoperability at the semantic level is represented by ontologies, which offer a formal conceptualization of the data that can be shared, reused, and aligned [8].

Specific and specialized training of bachelor and master degree students will be the prime focus throughout the duration of this work.

Awareness seminars and special workshops for IT students to discover and explore the importance of this domain are already in process and progress.

Acceptance of this potential work for funding will speed up these efforts. We will be able to incorporate more graduate researchers. This field will be polished for international standard Ph.D. research in our University.

PLAN OF WORK:

Activities	**Researcher responsible**
A thorough investigation of relevant natural language ontologies Develop a standardized approach for the creation, editing, mapping, and reusability of ontologies in local languages.	Existing progress. Conduction of discussion sessions, workshops, and seminars with relevant experts. Detailed Consultation with Software Engineering Local Researchers.
Social media and news are the specific domains for which we have to develop ontologies in Urdu (natural language). Collection and enrichment of domain specific knowledge and terminologies.	Explore the Web and other resources for relevant terms' discovery. Individuals use different word extraction, mining, and classification techniques.
To provide proof that searching, integration and mining of traditional web in local language is improved. Ensure that Novel contributions become a part of semantic web in our mentioned domain.	Collection of experimental evidence about our contribution to relevant knowledge creation, effective strategy about the effect of our developed ontologies.
A Web-based interface for Natural Language (NL) ontology development and query retrieval systems will be deployed.	Selection of the optimal Web Framework for the mentioned task and development of an international standard SPARQL-based query retrieval interface.

Like DODDLE-OWL for English and Japanese. Using any tool for natural language ontology in any domain is a tedious job [9].

The majority of editors work well with ontologies in English, Spanish, and French.

There is very little work done in Urdu or Pashtu. There is a lot of work done in Arabic and Persian. The creation of a natural language ontology for a particular subject is a crucial step in implementing the Semantic Web. There are several software tools

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2022 doi:10.20944/preprints202211.0201.v1

designed for creating domain ontologies of the most popular natural languages, but doing so for any particular natural language is difficult [10]. It's Urdu in natural language.

We applied the translation of the original text into English text and transformed the resulting English ontology to adopt this environment for any natural language (that has a dictionary on WordNet).

Many researchers propose a general procedure to construct domain ontology for any natural language using Protégé or any other standard ontology editor. Natural language ontology is also supported by protégé [13]. But a semi-automatic system for NLO development in protégé is a hot research area. Some researchers propose a general procedure to develop NLO in any ontology editor, principally protégé. Protégé has XSLT transformation support. Semi-automatic NLO development projects already working for Arabic and Persian languages. The text recognition or enrichment can be called as automatic.

A language called XSLT (eXtensible Stylesheet Language Transformations) is used to convert XML files into other XML files, or into other formats like plain text, HTML for web pages, or XSL Formatting Objects, which can then be turned into other formats like PDF, PostScript, and PNG[11]. Modern web browsers support XSLT 1.0 to a large extent. Ontology detected in a source language is translated into the target language using XSLT. The Protégé ontology editor's backup XML and RDF code is as follows.

<xsl:template match="rdfs:label">
<xsl:variable name="word" select="." />
<rdfs:label xml:lang="{ $sourceLanguage } ">
<xsl:value-of select="$word"/>
< rdfs:label xml:lang="{$targetLanguage} ">
<xsl:value-of select="$dict/word[@name=$word]"/>
< /rdfs:label>
</xsl:template>

If we run the Reasoner (Default) after editing, it will save as the given people ontology. The given node (edited) عورت is mentioned in the ontology as a literal of Female IRI.

#female عورتUrdu originally "ar" as a source language. The use of "ur" is under research and discussion.
< /owl:Class >

The Unicode Standard, Version 12.0 consists of Arabic and Persian alphabetic symbols. All the mathematical, old and new symbols are available from this link. **https://www.unicode.org/charts/.**

For Example1EE00 ﺍ) (ARABIC MATHEMATICAL ALEF→
FE8D arabic letter alef isolated form≈ < font> 0622 arabic letter alef.

The above Unicode Standard is not available for Urdu, an open research issue. The Traslated Ontology is shown in the below figure 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2022 doi:10.20944/preprints202211.0201.v1

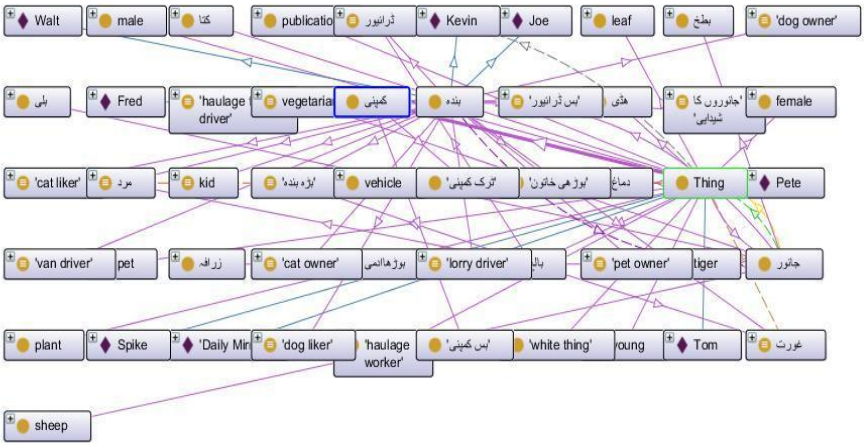

Fig 3. Another Expanded view of Translated People Ontology.

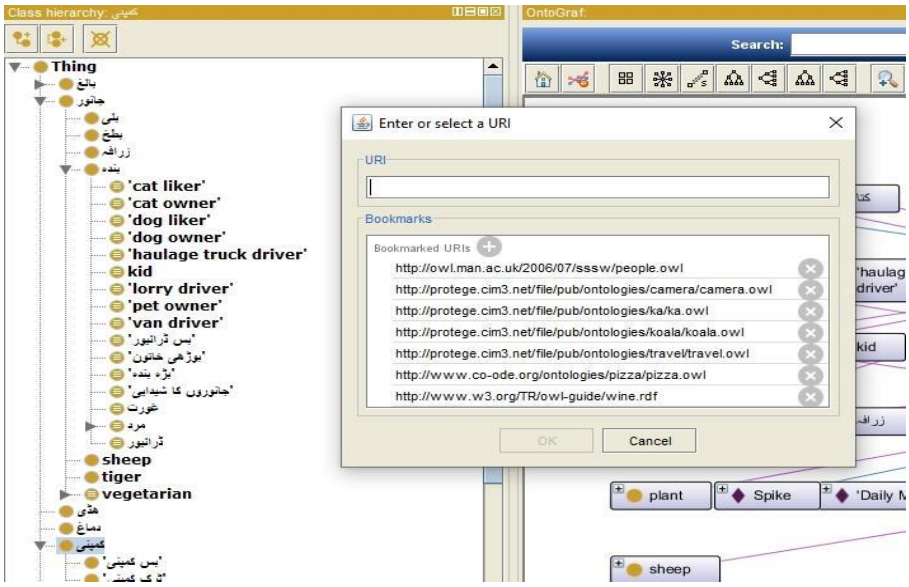

Fig 4. People Ontology Repository or Resource in Protégé Work Space

Up to now, XLST has been used, so there is a need for specific translators for Semantic Linked Databases. Building NLO for Arabic is robust because its complete Unicode is available. Partial Unicode of Persian is also available from the link below.

https://www.unicode.org/charts/. The area of Urdu or any other oriental language Unicode development is an open research area. Only UTF-8 and Web3 or Semantic Web are supported. Figure 4 show the Protégé work space.

References

1. Muhammad Ishaq, Abdullah Khan, Muhammad Asim, Javed Iqbal Bangash, Asfandyar Khan: Comprehensive Selective Improvements in Agri-Informatics Semantics. Journal of Information Science vol49 issue 3, (2022).
2. M. Ishaq, A. Khan, M. Khan and M. Imran, "Current Trends and Ongoing Progress in the Computational Alignment of Biological Sequences," in IEEE Access, vol. 7, pp. 68380-68391, 2019, doi: 10.1109/ACCESS.2019.2916154.
3. Liyang Yu. A Developer's Guide to the Semantic Web.. XXV, 829. https://doi.org/10.1007/978-3-662-43796-4, Springer Berlin, Heidelberg (2014).
4. Konstantinos G, Georgia T, Christos P et al. Associating ω-automata to path queries on Webs of Linked Data. Engineering Applications of Artificial Intelligence, Volume 51, 2016, Pages 115-123, doi.org/10.1016/j.engappai.2016.01.013.
5. Annarita Orsini, Alessio Innocenti. Semantic Web, ontologies and GIS for the cultural routes. *Netcom*, 32-3/4 | 2018, 377-384.
6. Sadeghineko, F., Kumar, B., Chan, W. . A Semantic Web-Based Approach for Generating Parametric Models Using RDF. Advanced Computing Strategies for Engineering. EG-ICE 2018. Lecture Notes in Computer Science (), vol 10864. Springer, Cham. https://doi.org/10.1007/978-3-319-91638-5_20 2018.
7. Mazayev, A., Martins, J.A., Correia, N. Semantically Enriched Hypermedia APIs for Next Generation IoT. Interoperability, Safety and Security in IoT. InterIoT SaSeIoT 2017 2017. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 242. Springer, Cham. https://doi.org/10.1007/978-3-319-93797-7_3 2017.
8. .Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).
9. Mael Jullien, Marco Valentino, André Freitas et al. Do Transformers Encode a Foundational Ontology? Probing Abstract Classes in Natural Language. diap Research Institute, Switzerland pre print 2022.
10. Jay Selig.Understanding Ontology and How It Adds Value to NLU. Expert.ai. https://www.expert.ai/blog/how_ontology_works_and_adds_value_to_nlu/. 25 August 2021.
11. Ali A, Mélanie A, Jérôme R, Ontology-based NLP information extraction to enrich nanomaterial environmental exposure database, Procedia Computer Science, Volume 176,2020,Pages 360-369,doi.org/10.1016/j.procs.2020.08.037.
12. Laukaitis, A.; Ostašius, E. Plikynas, D. Deep Semantic Parsing with Upper Ontologies. Appl. Sci. 2021, 11, 9423. https://doi.org/10.3390/app11209423.
13. Musen, M.A. The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003.
14. Alatrish E S,Tošić D, Milenković N. Building ontologies for different natural languages. Computer Science and Information SystemsVol 11(2) PP 623-644. 2014.