International Journal of Information Technology and Knowledge Management
July-December 2010, Volume 2, No. 2, pp. 351-354

A SURVEY: ISSUES OF SEMANTIC MATCHING FOR INDIAN LANGUAGES USING ONTOLOGY

S.M.Chaware[1] & Srikantha Rao[2]

Semantic matching is an important issue when two domains or contexts are to be compared. This also has prime importance when these domains are in different languages or scripts. There are many approaches for semantic matching like using relational operators, using dictionaries, using translations etc. All the approaches have some pros and cons and will not suite for any domain or specially for semantic matching within languages or scripts. In this paper, a survey has been made in semantic matching techniques using ontology, its, issues and mapping for Indian Languages.

Keywords: Semantic Matching, Ontology, Inter-Lingual-Index (ILI), Multilingual Dictionary, WordNet.

1. Introduction

Semantic matching is a technique used in computer science to identify information which is semantically related. A matching process for documents that uses knowledge of meaning to broaden recall. For example, semantic matching may use knowledge of synonyms. Given any two graph-like structures, e.g. classifications, database or XML schemas and ontologies, matching is an operator which identifies those nodes in the two structures which semantically correspond to one another.

For example, applied to two keywords, it can identify that a keyword labeled “car” is semantically equivalent to another keyword “automobile” because they are synonyms in English. This information can be taken from a linguistic resource like WordNet.

There are many semantic search engines. Some of the issues regarding these search engines are, first, the knowledge structures cannot be frequently undated, the context ontologies cannot be customized and updated, second, some engines has low precision and high recall values and third, these engines are being used for web pages designed and implemented for English language only, may not be suitable for Indian Languages.

The main motivation of this paper is to study existing semantic matching techniques using ontology, its issues and mapping them for Indian Languages.

[1]D.J. Sanghvi College of Engineering, Mumbai-56, India

[2]Late B. Hiray S.S.Trust’s Institute of Computer Application, Mumbai-51, India

Email: [1]smchaware@gmail.com, [2]dr_s_rao@yahoo.com

2. Semantic Matching Techniques

Semantic matching represents a fundamental technique in many applications in areas such as resource discovery, data integration, data migration, query translation, peer–to-peer networks, agent communication, schema and ontology merging. Semantic similarity refers to semantic closeness or nearness. There are 3 types of semantic similarity: surface, structure and thematic [3]. Surface, structure focuses on concepts respectively and relationships and thematic is based on both.

Semantic matching techniques may include use of search engines that are more likely to understand the meanings of hidden in retrieved documents and user’s queries by means of adding semantic tags into texts. The semantic matching primarily concentrates on semantic search engines, semantic search methods, hybrid semantic search engines, XML search engines, ontology search engines, and semantic multi-media search engines [7].

Each search engines has its own technique and method. Semantic search engines uses iteratively cyclic mechanism, which generates two classes based on exact search, whereas XML search engines are based on full-text search, which sees XML documents as a collection of structured texts, and then executes a series of algebra query language to retrieve [7]. Ontology search engines are designed for querying ontological files. In this, ontology registry is designed to store metadata about ontologies and ontology server stores the ontology [7]. Ontology search in ontology registry is executed by either query-by-example or query-by-term. Some of the examples of are Swoogle and OntoKhoj [7].

3. Introduction to Ontology

Ontology can be defined as “An explicit specification of a conceptualization” [1]. Ontology is a conceptual

representation of the entities, events and their relationships. Two main relationships are abstraction and composition (part-of). Ontology is arranged in a lattice or taxonomy of concepts in classes and subclasses. Each concept is typically associated with various properties describing its features and attributes as well as various restrictions on them. It is a shared conceptualization of knowledge in a particular domain. We distinguish top-level ontologies describe very general concepts like space, time, matter, object, event, action etc. which are independent of a particular problem or domain [1]. Other ontologies are domain and task related to domain or activity.

Examples of existing ontologies are:

- Top-level ontologies are SUO (Standard upper Ontology) provides definition for general purpose terms.
- SENSUS: natural language-based ontology developed by NLG at ISI to provide a broad conceptual structure for working in machine translation.
- WordNet: is a large lexical database for English created at Princeton University or IITB.
- Medical ontologies: gene, Galen and Menelas.

4. LEVELS OF ONTOLOGY FOR SEMANTIC MATCHING

Ontology must have at least four levels, two language independent levels: conceptual and meta-conceptual and two language dependent levels: linguistic level and instance [4]. These levels are briefly described below.

1. The Conceptual Level: It is organized around the pivot which is represented by the unique ID. Conceptual names do not correspond directly to the language referents; they enable a definition of some relations on the conceptual level, such as synonym, for example, France and French, meronymy, for example, Paris => France => Europe, and predication, for example, Paris is the capital of France [4]. The relationships are established with WordNets for which the concepts of Inter-Link-Index (ILI) were introduced.

   The conceptual level is so important, which organizes information objects into groups or categories, where each category represents a relevant concept interpreted in the problem domain. The main types are:

   - Conceptual Taxonomy: It is a hierarchical organization of concept descriptions according to generalization relationship. Each concept has alike to its super-concepts and sub-concepts. Generally it is created manually [1].
   - Formal or Domain Ontology: Ontology can serve as is source description and can be used for query formulation. Conceptual graphs obtained from queries have been linked to ontology by using lexical conceptual graphs. Ontology can be represented by lattice as formal basis and proposed for Ontoquery project. [1].
   - Thesaurus: It is a collection of words or phrases linked through a set of relationships including synonymy, antonymy and 'is-a' relationship [1].

   Features of Conceptual level:

   - Conceptual structures may be concept taxonomy, domain ontology, top ontology, linguistic ontology, semantic linguistic network, predictive and dictionary.
   - Representation of a conceptual structure can be using tree, semantic network, context vectors, conceptual graphs etc.
   - Relationships supported by a conceptual structure can be subsumption, a kind-of, a part-of, associations etc.
   - Conceptual Structure can be created by manually, automatic learning methods or using NLP [1].

2. The Meta-conceptual Level: This level enables a homogenous classification of proper names on the bases of super-type and type that are associated to every proper name, where super-type classifies proper names according to their traditional syntactic and semantic properties. Super-types proper names can be distinguish as historical, relious and fictious names [4].
3. Linguistic Level: This level describes the realization of a proper name in the observed language. On this level the canonic forms or prolexemes are defined and are connected to the ID for the particular language. The aliases are connected to prolexemes that describe the variations in orthography, abbreviated forms, acronyms etc [4].
4. The Level of Instances: This level contains the inflected forms of proper names that are linguistically described [4].

5. ONTOLOGY FOR SEMANTIC MATCHING

Many search engines works semantically by analyzing the context of words in their index and returning likely matches for the same. These search engines are knowledge-based. They are using ontologies as knowledge to understand the meanings of the concepts. Ontologies for semantic matching are preferred due to many reasons. First, ontology defines

well-formed structure of the domain with all relationships between the entities. Second, it is easy to form a query and traverse along with the ontology to find the match. Synonym and polysemy are two issues. Synonym is a word that means the same as another word and polysemy means many words with multiple related meanings.

6. Ontology Development Tools

To build ontology for a huge amount of data, number of tools are available such as text2Onto, the ASIUM system, the Mo'k Workbench, terminae, OntoLearn or OntoLT, OntoEdit, OntoBroker, OntologyBuilder and Ontology Server, KAON, Jena2. But these tools are being used to build ontology from a given set of textual data. Only KAON and Jena 2 supports storing of ontology using RDBMS.

These tools have many drawbacks such as:

It does not exist a detailed methodology or method that guides the ontology learning process from text.

- The methods are mainly based on natural language analysis techniques, and use corpus that guide the overall process.
- Work uses domain and general corpora to remove unspecific domain concepts from an existing ontology.
- The most common ontology used by many methods is WordNet, which is used as initial ontology enriched with new concepts or relations..
- All these methods require the participation of an ontologist to evaluate the final ontology and the accuracy of the learning process.

7. Issues of Semantic Matching using Ontologies

Some of the issues of semantic matching using ontologies are as follows:

- Ontology Building: There are many entities that should be kept in mind while building ontology. Some are ontology area and domain, ontology resources, keywords in ontology, attributes and its values of the classes etc [6]. Ontology can be build either manually, semi-automatically or automatically.
- Ontologies Matching: There are number of approaches for matching ontologies. Most of these are based on lexical techniques. For example, PROMPT, GLUE, QOM and IF-Map etc. But these methods can not be applicable in certain domain; new algorithms are to be developed.
- Matching through Mapping: Matching of ontologies can be done by mapping methods. Mapping from one synset to another synset can be done by using lexical transfer engine, like WordNet for each language. But, the problems may occur when ontologies belong to a particular domain. There can be more than 4 situations occurs for mapping like one-to-one, many-to-one, one-to-many and no link.
- Matching Approaches: There can be many matching approaches within the synsets like individual and combined, hybrid and composite, schema and instance, element and structure i.e. schema-based.
- Mapping Evaluation: Mapping can be evaluated by using proper algorithms, criteria, measuring instruments and methodology like set of tools and methods.

8. Semantic Matching using Ontology for Indian Languages

Indian languages consist of words, called as synsets. These synsets can be defined with their meaning in a WordNet for each language. These WordNets are readily available for Hindi and Marathi languages. Synsets are with respect to noun, verbs, adjectives, adverbs and relationships. Nouns can be used as keyword for conceptual domains. Semantically equivalent synsets between WordNets of different languages are interlinked using Inter-Lingual-Index (ILI) links. These links may be one-to-one or one-to-many or many-to-one etc. Mapping of ontologies for semantically same synsets can be achieved through matching algorithm.

Ontology can be build by using RDBMS package such as Oracle [2], MS-SQL or by designing and implementing graph-like structure or by using ontology languages like RDF, XML etc. Synsets can be retrieved from each WordNet, are to be stored in a dictionary called as Multilingual Dictionary [5]. Semantically matching can be achieved through mapping algorithm, by considering all situations. Exact keywords which are semantically same, can be found by either alphabetically first word from synset or level of keyword from ontology.

9. Conclusion and Future Work

In this paper, we made a survey of semantic matching techniques, issues of using ontology for matching and solutions to those issues to match semantically for Indian languages using ontology. We found that, these issues can be addressed by using existing ontology methods, tools and semantic matching techniques with proper modification to suite for Indian languages. The future scope is to develop a system which will work efficiently for semantically matching using ontology.

REFERENCES

[1] O. Dridi, Riadi LAB, "Ontology-based IR".

[2] S.Das, E.I. Chong, G. Eadon, J. Srinivasan, "Supproting Ontology-based Semantic Matching in RDBMS".

[3] J. Mustafa, S. khan and Khalid Latif, "Ontology-based Semantic IR".

[4] C. Krstev, D. Vitas, D. Maurel, M. Tran, "Multilingual Ontology of Proper Names".

[5] Dr. Pushpak Bhattacharyya et al, "Synset Based Multilingual Dictionary: Insights Applications and Challenges".

[6] Xiaohuan Zhang, Wenjie Li, "Ontology-based Retrieval System".

[7] Hai Dong, F.K.Hussain, Elizabeth Chang, "A Survey in Semantic Search Technologies".