CHAPTER 4

SEMANTIC MATCHING APPROACHES

4.1 INTRODUCTION

Semantic matching is a technique used in computer science to identify information which is semantically related. In order to broaden recall, a matching process for documents uses knowledge. For example, semantic matching may use knowledge of synonyms. Given any two graph-like structures e.g. classifications, database or XML schemas and ontology, 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 be identified 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 [6].

There are many semantic search engines. Some of the issues regarding these search engines are, first, the knowledge structures cannot be frequently updated and the context ontology 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 is to study existing semantic matching techniques using ontology, its issues and mapping them for Indian languages like Hindi and Marathi.

4.2 THEORETICAL FOUNDATION

In order to match two strings semantically, we need some background information to build a system that will work efficiently. In this section, we focus on some of the terms needed for semantic matching as follows:

4.2.1 ONTOLOGY

There is no formal definition of ontology. The definitions can be categorized into three groups.
1. Ontology can be defined as ‘Theory of existence” in philosophy.
2. Ontology is an explicit specification of conceptualization.
3. Ontology is a body of knowledge describing some domain, typically common sense knowledge domain.

The second definition has been accepted widely comes from Thomos Gruber [7]. Explicit specification of conceptualization means that ontology is a description of the concepts and relationships that can exist for a domain. A conceptualization can be defined as an intentional semantic structure that encodes implicit knowledge constraining the structure of a piece of a domain. Ontology is a specification of this structure; i.e. it is a logical theory that expresses the conceptualization explicitly in some language. Conceptualization is language independent, while ontology is language dependent. Ontology is an important for the purpose of enabling knowledge sharing and reuse. Practically ontology is an agreement to use a vocabulary in a way that is consistent with respect to theory specified by ontology [8].

The representation of a body of knowledge is based on the specification of conceptualization. A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose. There exists a set of objects called universe of discourse. Ontology is used to represent the entire set of objects in a prescribed way to represent knowledge. The backbone of ontology is taxonomy. Taxonomy is a classification of things in a hierarchical form. It is usually a tree or a lattice that express subsumption relation [9].

Natalya F. Noy and Deborah L. McGuinness [45] in their paper have described simple rules in ontology design which can be referred in many domains. They can be used to make design decisions in many cases. The fundamental rules are:

- A domain cannot be modeled by a single way, but the solution depends on application of that domain and its extension.
- Ontology development is an iterative process.
- Concepts in ontology should be closed to the physical or logical objects and their relationships should be in domain of interest. The domain can be described by nouns and verbs in sentences that described domain.
4.2.2 NEED OF ONTOLOGY

Ontology gives the exact representation of knowledge, which can be used for semantic matching. Xiaohuan Zhang and Wenjie Li in their literature [47] have described the various reasons of using ontology for semantic retrieval in web. Some are, first, no structured information for web, ambiguity of web content, inadequacy for automatic information transfers, inexistence of universal mechanisms and so on. The same reason can be applicable to use of ontology for semantic matching.

In general, following are the reasons to use ontology.

- Ontology that describes the domain is a particular slice of reality, for example, company’s operations, pizza restaurant and so on.
- When we need a detailed model or picture or schema of a slice of reality which is based on the facts that we know about that reality then we need ontology. Ontology will describe all the details about the domain along with knowledge that may be used for assertion.
- Ontology gives an effective knowledge representation, which may be useful for performing analysis of the field.
- Ontology enables reuse and knowledge sharing.

4.2.3 FORMAL REPRESENTATION OF ONTOLOGY

Ontology is to be used for automatic processing in computers, need to be specified formally. There are several languages that are used for expressing ontology. The figure 4.1 shows the levels of expressing ontology [57]. On the left side, the terms are defined to express the knowledge only. The ontology with formal specification of conceptualization is when at least one formal relation is defined and used between terms. This relation is expressed in formal language and it is possible to conclude implications that follow from the fact that two terms are connected with the relation. In all form, the terms are related by some relationships between each other which form a hierarchical structure to represent knowledge.
Figure 4.1: Various Forms of Ontology

Figure 4.2 shows the graphical representation of ontology with two concepts C1 and C2 are related with relation R. The concepts and the relation can be defined as some lexical references [52].

Concept C1  Concept C2

Relation R

Figure 4.2: Graphical Depiction of Ontology

Semantic network also called as concept network is a graph, where vertices represent concepts and edges represent relations between concepts. The possible relations between concepts that are used are synonym, antonym, meronym, hyponym [10]. Conceptual graphs are a logical formalism that includes classes, relations, individuals and quantifiers. It has direct translation to the language of first order predicate logic, from which it takes its semantics.
4.2.4 LIFECYCLE OF ONTOLOGY

Natalya F. Noy and Deborah L. McGuinness [45] in their paper described some rules that define the complete lifecycle of the ontology. The rules can be used as guidelines to define and develop ontology for any domain, which is used to make decision in any environment.

The rules can be defined as follows.

- The domain and scope of the ontology can be determined by using some competency questions.
- Reuse any ontology exist for the domain.
- Enumerate important terms in the ontology that describe the domain.
- Define the concepts and sub-concepts as classes and its inheritances.
- Define the properties of the classes as slots that describe the concept.
- Define the facet of the slots.
- Create an instance of the classes that describe each object of the domain.

If we follow some of the rules but in sequence, we can develop ontology for any domain.

4.2.5 STORING AND QUERYING ONTOLOGIES

The ontology can be stored in relational database systems as directed tree or directed acyclic graph. In both structures, a path is a sequence of nodes that are connected by directed edges. The length of the path is the number of nodes it contains. In a graph, we can compute its successors or ancestors by traversing the graph forward or backward. In relational database systems, the ontology can be stored as node-table and edge-table separately. Typical queries are classified as query to find reachability, query to find successor or ancestor and so on. The queries can be extended to support in-order or post-order form from the tree structure of ontology [50].
Sourpriya Das et al. [51] had addressed the problems of supporting ontology-based semantic matching in RDBMS. They proposed the solutions with two new operators along with new indexing scheme to speed up the matching operations. Ontology has been stored in OWL format as system-defined tables.

4.2.6 SYNSETS FOR SEMANTIC MATCHING

Synsets define the many alternative words for a single word within a language. It also represents a meaning for a word. This meaning can be represented by a single word or a phrase. Synsets can be extracted from a WordNet for a specified language. In semantic matching, we have either to translate a string from one language to other or use WordNet. Using WordNet for Indian languages will not be feasible solutions as it leads to a complex system or WordNet are not available for many Indian languages. Instead, we can maintain domain-specific synsets for each language as multilingual dictionary, as suggested by Dr. Puspak B. in his literature [11]. This will help in identifying the correct word that matches semantically with the entered word. So, there is no need to search an entire WordNet for each language.

4.2.7 SOME COMMON ISSUES

Many semantic search engines along with methodologies had been proposed which uses ontology as a main driving force behind semantic matching. But Hai Dong et al. [53] in their paper identified some common issues related to both. They are as follows:

- Differentiation between designers and user’s perceptions for semantic matching, ontology terms, relations, domain’s objects perceptions and so on.
- In real scenarios, the domain specifications remain static in terms of knowledge structure. The semantic search engines or ontology structure cannot be frequently updated to suite the change of user’s requirements and external environment.
- Some semantic search engines or approaches may not show performance in improving precision and lowering recall values.
4.3 ON-HAND APPROACHES

4.3.1 ONTOLOGY SUPPORTED SEMANTIC MATCHING FOR DOCUMENTS

4.3.1.1 Approach 1

Morteza Poyan rad et al. [12] proposed query expansion in their approach by considering all kinds of relationships. The approach had proposed for English and Persian documents. Ontology is being used to extract the concepts from the documents and queries. According to the query terms, it is translated to either language. Phrase translation is being used instead of word translation once it is recognized. Query extension may be before or after translation.

The steps for Cross-Language searching are:

- Extract all synsets from WordNet for entered word.
- If for any word more than one synset exists then calculate semantic similarity for this word along with nearby words.
- Select most related synsets for the entered word based on semantic similarity.
- All the synonyms, all the hyponyms and one hypernym are added for the entered word.
Figure 4.3 shows semantic matching approach for documents, in which semantic matching has been done by computing the concepts from one language documents, forming ontology from concepts and translating those concepts into target language documents. The semantic similarity of two words from WordNet has been calculated from information content (IC) of concept as: IC (Concept) = - \log (P (concept)), where P (Concept) = 1/number of hyponym for concept. Bi-lingual dictionary is being used for translation of English queries to their equivalents Persian. If the phrase does not exist in a dictionary then it will be broken into words and stage wise combinations of words are being tested in a dictionary for translation. This way English document is retrieved with semantic matching.

### 4.3.1.2 Approach 2

Atanas K. Kiryakov and Kritil Simov [13] in their paper have suggested methodology based on concept indexing and query expansion for documents to match semantically. In order to select the most relevant documents based on user’s query, ontology has been built based on the concepts represented in the text and in the query. A document is converted into a bag of words, stems or other textual elements called as atomic text entities (ATE). These elements define the concepts represented by chunks of text also define the topics of the document. These elements are verified with the elements of the query to match. Figure 4.4 shows the approach, where the real challenge is to find ATE from the documents.

![Figure 4.4: Atanas Kiryakov Semantic Matching Approach](image)
4.3.2 DICTIONARY-BASED SEMANTIC MATCHING

4.3.2.1 Approach 1

There are many disadvantages of bilingual dictionaries given in [11]. First, it is possible that one sense of \( W_i \) in \( L_1 \) is exactly as one of the senses of \( W_j \) in \( L_1 \). Second, there is no mechanism to relate \( W_i \) with \( W_j \) in \( L_1 \), though they express same meaning. Here, \( W_i \) is the word in \( L_1 \) and \( W_j \) is another word in the same language \( L_1 \). Third, there may be duplication of manual labor work to develop many bi-lingual dictionaries [11]. If machine translation (MT) system is to be developed involving \( n \) languages, we have to create \( n^*\frac{(n-1)}{2} \) language pair wise dictionaries. No major machine translator or cross-language information retrieval (CLIR) project involving multiple languages has ever used concept-based dictionaries [11].

Pushpak Bhattacharyya et al. [11] has suggested synset-based multilingual dictionary for \( n \) languages, in which synsets of multiple languages not words are linked among themselves or through inter-lingual indices. They developed multilingual dictionary for Hindi, Marathi and English based on concepts with synsets. The members of a particular synset are arranged in the order of their frequency of usage for the concept in question. Monolingual or bilingual or domain specific dictionary can be built from this multilingual dictionary. In order to link synsets of one language to another, four linking situations have been considered as one-to-one, one-to-many, many-to-one and none. They defined two ways to express the concept from one language to another, first by using a syntactic expression and second, by using transliteration.

Figure 4.5 shows the synset-based multilingual dictionary approach for semantic matching. The proposed approach has many advantages. First, semantic features are assigned to nominal concepts and not to any individual lexical item of any language. These features are stored only once for each row and become applicable for any language. Second, the dictionary can be used as monolingual or bi-lingual. Third, the language group which lacks competence in pivot, can benefit from already worked out languages. However, the challenges involved in maintaining concepts for each language are, first, a concept may be expressed using different syntactic category in different languages, second, a concept may be
expressed through a syntactic expression in one language or through a single word in another language.

![Diagram of information retrieval process](image)

Figure 4.5: Synset-Based Multilingual Dictionary for Semantic Matching

### 4.3.2.2 Approach 2

A. Kumaran [3] has given the semantic matching strategy in his PhD dissertation as follows. Convert the query string to a set of concepts using standard linguistic WordNet. The matching of two different words is based on concept called semantic closure computed using WordNet. Resulting semantic primitive has been compared for equivalence, generalization or specialization based on common concept hierarchy between the languages. In order to find the semantic matching between two local language strings, compute transitive closure has computed by a function using ‘IS-A’ relation within a language and ILI across languages. Entered query string has been checked for match with WordNet of target language. Also each concept and its hypernyms have been verified with ILI for each synset. If it matches then add all the members from WordNet of target language to a matched set. Figure 4.6 shows two ontologies matching as semantic matching.

Inter lingual index (ILI) that describes the equivalence relations between synsets in different languages. ILI links two WordNets, where synsets for every individual WordNet which relate concepts on the basis of topic has been defined with the hierarchy of domain labels. Two ways to build ILI has been mentioned in the literature [46], first, each language synset can be built from WordNet by using simple cross-language synonym links. The second option is to maintain ILI links as unstructured links to WordNet by using cross-language hypernym links.
4.3.3 GENERAL CATEGORY FOR SEMANTIC MATCHING

Stuart Aitken and Sandy Reid [14] in their paper had implemented a system which worked on ontology and compared with simple keyword techniques for information retrieval. The retrieval tool operated on both keyword and ontology-based word in order to match query words with records, both of which consist of free text. The lexical terms which are associated with concepts are used to extract ontology concepts from the text fields of records. Query terms are being matched with concepts and more terms are been added with AND, OR connectives. In this case, sub-concepts of query concepts which occurred in cases are considered as matching the query. The query has been expanded to include sub-concepts. Different queries have been tested against ontology-based and keyword-based systems for different data sets. The performance was calculated with the precision and recall values. The result showed that high precision comes from the use of unambiguous terms in keyword matching or in the lexical terms associated with concepts. There were variations in recall values due to tokenizing of the input and use of unknown terms in data.

Sandhya et al. [15] in their literature described different approaches based on query expansion that includes various combinations to form a query. Ontology has been built with possible concepts in the domain, identify properties and range as terms for ontology. In their query expansion module, the query has been formed by taking either concept or property or
instance or various combinations of concept, property, instance or a pair of instance etc. With respective to these query terms, ontology has been traversed and searched for information retrieval.

Ashish Sureka et al. [16] has implemented a domain specific ontology from common sense semantic network. They took product domain as camera and phone and matched the words semantically from both the ontology. The terms for ontology has been extracted from various sources like product feature extraction, opinion and sentiment expressing phrase extraction, semantic orientation determination of sentiment expressing phrases and by identifying the intensity or strength of sentiment expressing phrases. The concepts from these sources are related by some relations such as IS-A, MadeOf, UsedFor, DesireOf etc. Customer’s feedback or inputting product features are being matched to appropriate node in the product domain ontology and unmatched has been omitted. The evaluation has been made with the customer’s product features and company’s product feature to match. Figure 4.7 shows the system architecture for semantic matching using ontology for product domain.

![Figure 4.7: Semantic Matching for Product Domain Using Ontology](image)

Ontology matching has been performed where information sharing and reuse becomes necessary in ontology development. In ontology matching, lexical similarity is performed using synset from WordNet. JungAe Kwak et al. [17] has suggested a super word set, in which hypernym, hyponym, holonym, and meronym set were defined similar to WordNet for
matching. Similarity was calculated by the rate of words of concept name and synsets words inclusion in the super word set. The methodology has been proposed for matching of concepts, matching of properties and property unmatched concepts from the result of ontology matching. Three levels of similarities like between concepts, between properties and logical inference measure were mentioned. The similarity measurement in ontology matching is taken by considering lexical, structural, instance and logical inference. The ontology matching process has been defined as starts with concepts then between properties and between logical inferences. The performance of the system has been measured in terms of precision and recall for ontology matching through the proposed method and manual matching.

Alfio Ferrara et al. [18] have explored the use of ontology for providing the resources to the other nodes in open distributed environment in their research project on ‘Methods and techniques for Ontology Matching and Evolution in Open Distributed Systems’. Ontology is used for describing the data and resources. Entire environment is distributed into two layers, first, content knowledge layer and second, network knowledge layer. In each layer, a content concept is characterized by a set of properties and a set of semantic relations with other concepts. Matchmaking techniques were provided for supporting the semantic retrieval of information and resources over open distributed systems. Also the evolution techniques were considered for supporting the registration of new knowledge acquired by other nodes through node interactions. The final goal was to find concepts that have a semantic affinity with a target concept. Three functions were defined as, affinity function \( A(t, t') \rightarrow [0,1] \), data type compatibility function \( T(dt, dt') \rightarrow \{0,1\} \), and property and relation closeness function \( C(e, e') \rightarrow [0,1] \). Also three relations are defined among the concepts of two ontologies. First, SYN, where each synset of concepts of one ontology will have both hypernymy and hyponymy or metonymy relations with the ontology’s concepts. Second, BT/NT, this was more general or broader relations between the concepts, and third was RT defined as part-of or specification of both the concepts from two ontologies. Each relation was defined with some weight in order to match. Finally, the relationship of those weighted relations is defined as \( W_{\text{SYN}} \geq W_{\text{BT/NT}} \geq W_{\text{RT}} \), where W are the weight of each relation.
4.3.4 DRAWBACKS OF ON-HAND APPROACHES

We found some drawbacks of on-hand approaches. They are as follows:

- Semantic similarity between the concepts of the documents may vary since it depends on number of hyponym of the concepts in a document.
- ATEs from documents cannot be computed easily since it may depend on one’s perception for a string in a language.
- Maintaining machine readable dictionary (MRD) will be challenge. Also, representation of concepts in different languages may differ.
- Most of the approaches use synsets from WordNet for each language. It may be time consuming to find out the synset from WordNet.
- Some of the synsets are not found in WordNet which are related to particular domain.
- ILI has to be exclusively maintained and used for semantic matching as suggested by A. Kumaran in his PhD thesis. Maintaining ILI across the languages will increase the cost in terms of storage and processing.
- In one of the approach, the semantic matching has been done with respect to customer’s expected product features and company’s product features. But there may be an ambiguity in representing the features.

4.4 CLASSIFICATION OF SEMANTIC MATCHING APPROACHES

4.4.1 MACHINE TRANSLATION (MT) APPROACH

In MT approach, we use software in order to translate text or speech from one natural language to another. MT performs this by applying simple substitution of words from another language. Current MT software allows for customization by domain or profession, improving output by limiting scope of allowable substitutions. MT uses a method based on linguistics rules, based on transfer or inter-lingual or dictionary. These rules can be applied for either direct translation or intermediate translation.
In CLIR, MT can be implemented in two ways. In the first approach, source language is translated to English by MT system and then the query terms are translated to target language by MT system. So, we have to use the ‘triangulation’ process. MT performs simple substitution of words from one language into another language. The most suitable words of the target language will replace the ones in source language. In the second approach, using MT system the source language query terms are being translated to target language terms directly and then retrieved by classical IR techniques. An ambiguity problem exists in the MT system since the translated query does not necessarily represent the sense of the original query. MT systems normally attempt to determine the correct word sense for translation by using context analysis [19].

4.4.2 DICTIONARY-BASED QUERY TRANSLATION APPROACH

In dictionary-based query translation approach the query terms can be a single word or a phrase. So, we need a combination of phrase reorganization, pattern-based phrase translation and query expansion before and after translation to improve dictionary-based query translation [49].

In dictionary-based query translation approach, query terms are translated into target language terms by using Machine Readable Dictionaries (MRD). MRDs are electronic form of various kinds of dictionaries such as general, domain specific, or combination of both. This approach is easy as compare to MT approach, but the problems of ambiguity in terms of word inflection, translation of compound words, phrases, proper names, spelling variants need to be handled. Some solutions have been suggested in order to handle these issues such as speech tagging, corpus-based disambiguation methods, query structuring and so on [21,22].

4.4.3 CORPUS-BASED APPROACH

A corpus is a repository of collection of natural language materials, such as text, paragraphs and sentences from one or more languages. Two types of corpora have been used in query translation, first, parallel corpora, in which the same text is represented in more than one
language. An alignment will show which sentence of the source language corresponds to exactly which sentence of the target text. When retrieving with parallel corpora, no need to translate the query from source language to target language since a source language query can be matched against the source language component of the corpus, and then can be aligned to the target language components for retrieval. The alignment process can be done by comparing both the language texts by the presence of indicators. The indicator could be an author name, date, source, special names, number or acronyms [19].

The second type of corpora is comparable corpora, which contain text in more than one language. The texts in each language are not to be translated to each other, but cover the same topic area, and hence contain an equivalent vocabulary.

4.4.4 MATCHING STRATEGY

The classification mentioned in above section leads to some building blocks to be used for matching system. The following aspects of a matching system have to be taken into account.

- How to translate text string from one language into another language? Are there any direct methods available?
- How to find an equivalent text string entered by a user, those can be either from same language or from different language?
- How to represent data into knowledge that may lead from analysis to decision?
- How to generate domain dictionaries and use of those dictionaries in order to solve ambiguity problems in matching, spelling variants, word inflection and so on?

4.4.5 DRAWBACKS OF EXISTING MATCHING STRATEGY

- There may be translation ambiguity since the translated query does not necessarily represent the same sense of the original query.
- There may be ambiguity in terms of word inflection, translation of compound words, phrases, proper names, spelling variants and so on.
- There may be ambiguity in representation of words in MRD.
In corpus-based approach, there are issues related to use of corpora and their alignment.

4.5 ONTOLOGY FOR MATCHING

4.5.1 ONTOLOGY BUILDING APPROACHES

4.5.1.1 History

Basically, a series of approaches have been reported for developing ontology. In 1990, Lenat and Guha published the general steps and some interesting points about the Cyc development. Initially, the Enterprise Ontology and the TOVE (TOronto Virtual Enterprise) project ontology had been proposed in the domain of enterprise modeling. Bernaras et al. presented a method used to build ontology in the domain of electrical networks as part of the Esprit KACTUS project. The methodology METHONTOLOGY appeared at the same time. In 1997, a new method was proposed for building ontology based on the SENSUS ontology. Some years later, the On-To-Knowledge methodology appeared as a result of the project with the same name. However, all these methods and methodologies do not consider collaborative and distributed construction of ontology [21]. In this section, we are describing some of the methodologies for building ontology.

4.5.1.2 Skeleton Methodology

This methodology is based on the experience of developing the Enterprise Ontology i.e. ontology for enterprise modeling processes. A plan or draft for a project along with activities can be represented as ontology.

A comprehensive methodology for developing ontology need to include the following stages:

- Identify the main purpose of the ontology: The purpose can be stated with a document that may contain some competency questions relating to the domain, but few in numbers.
- Build the ontology: The key concepts and its relationships can be captured, code the ontology in any coding language such as Prolog, Conceptual Graphs, L-Lilog, Ontolingua and so on and then integrate with any existing ontology.
• Evaluation of the ontology: The evaluation can be based on knowledge sharing technologies.
• Documentation of ontology: For effective knowledge sharing, adequate documentation of existing knowledge bases and ontology are needed. This documentation may include the main concepts defined in the ontology primitives to express the relationships in the ontology.

In addition to this, it should include a set of techniques, methods, principles and guidelines for each stage, as well as indicating what relationships exist between the stages. We can use either top-down or bottom-up approach to represent the ontology [23, 24]. This method is simple to implement but limited in scope.

4.5.1.3 Gruninger and Fox Methodology

Gruninger and Fox [22] proposed a methodology that is inspired on the development of knowledge-based systems using first order logic. This methodology has been suggested as TOVE project ontology within domain of business processes and activities modeling. This represents logical model of knowledge.

The steps are:

• Capture the motivating scenarios.
• Formulate informal competency questions, where the scope of the ontology can be decided.
• Formulate formal competency questions, which specify the terminology with definition and constraints.
• Specify the axioms and definition within the formal language.
• Specify the conditions under which the solutions to the questions are complete.

In this methodology, the ontology can be built by using questions and answers for motivating scenarios, which represents main concepts, properties, relations and axioms on the ontology [21][22]. The methodology may extend the scope but the procedure is complex.
4.5.1.4 Methontology Methodology

This methodology gives the construction of ontology at knowledge level. It has main activities from software development process and in knowledge engineering methodologies. This methodology includes the identification of the ontology development process, a life cycle based on evolving prototypes and techniques to carry out each activity in the management, development-oriented and support activities.

The ontology development process was:

- Determine the tasks to be performed when building ontology i.e. scheduling, control, quality assurance, specification, knowledge acquisition, conceptualization, integration, formalization, implementation, evaluation, maintenance, documentation and configuration management.
- Determine the life cycle of ontology as number of stages. This represents the activities to be performed in each stage and how the stages are related.
- Determine the techniques used in each activity, the evaluation of that activity for a product [21] [22].

These methodologies deal with software engineering concepts. It handles all the activities in detail. It also considers the activities performed during the development of ontology which may involve other activities from another ontology which is already built or under construction.

4.5.1.5 Sensus Methodology

This methodology had proposed to link domain specific terms to the huge ontology and to prune, in the huge ontology those terms that are irrelevant for the new ontology we wish to build. The result of this process is the skeleton of the new ontology, which is generated automatically using the steps given below. The method based on sensus is a top-down approach for deriving domain specific ontology from huge ontology [22].
The steps were:

- A series of terms were taken as seed i.e. key domain terms.
- These seed terms were linked by hand to SENSUS.
- All the concepts in the path from the seed terms to the root of SENSUS are included.
- Terms that could be relevant within the domain and have not yet appeared are added.
- Finally, for those nodes that have a large number of paths through them, the entire subtree under the node is sometimes added. The idea was that, if many of the nodes in a subtree have been found to be relevant then the other nodes in the subtree are likely to be relevant as well [22].

This methodology uses the existing ontology where the merging may be complex due to different structures.

4.5.1.6 WordNet Methodology

WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable and to support automatic text analysis and artificial intelligence applications [22].

The hypernym or hyponym relationships among the noun synsets can be interpreted as specialization relations between conceptual categories. In other words, WordNet can be interpreted and used as a lexical ontology in the computer science sense. However, such ontology should be corrected before being used since it contains hundreds of basic semantic inconsistencies such as (i) the existence of common specializations for exclusive categories (ii) redundancies in the specialization hierarchy. Furthermore, transforming WordNet into a lexical ontology usable for knowledge representation should normally also involve (i) distinguishing the specialization relations into subtypeOf and instanceOf relations (ii) associating intuitive unique identifiers to each category. WordNet has also been converted to a formal specification by means of a hybrid bottom-up top-down methodology to automatically extract association relations from WordNet. These associations have been
interpreted in terms of a set of conceptual relations, formally defined in the DOLCE foundational ontology [23].

Table 4.1: Comparison of Ontology Building Approaches

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Name of Ontology Building Approach</th>
<th>Methodology</th>
<th>Input</th>
<th>Specialization areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Skeleton Methodology</td>
<td>Top-down or bottom-up approach.</td>
<td>Plan or draft of a project.</td>
<td>Enterprise Ontology</td>
</tr>
<tr>
<td>2</td>
<td>Gruninger And Fox Methodology</td>
<td>Knowledge-based system.</td>
<td>Motivating scenarios.</td>
<td>TOVE Ontology</td>
</tr>
<tr>
<td>3</td>
<td>Methontology Methodology</td>
<td>Expand the features of software engineering aspects to knowledge.</td>
<td>Raw data as knowledge.</td>
<td>Chemicals Ontology</td>
</tr>
<tr>
<td>4</td>
<td>Sensus Methodology</td>
<td>Use of existing ontology; top-down approach.</td>
<td>Existing ontology.</td>
<td>Domain specific Ontology</td>
</tr>
</tbody>
</table>

4.5.2 DRAWBACKS OF EXISTING ONTOLOGY DEVELOPMENT APPROACHES

The following are some of the drawbacks in existing ontology development approaches.

- Some of the methodologies are too formal and only useful for small-scale applications or contexts.
- Some methodologies like Methontology, is more mature and detailed where as some steps can be either integrated or rejected depending on context [24].
- Integration of existing ontology may be difficult due to change in structure or plan.
- For each scenario, we cannot decide the competency questions which may represent the definition and constraints of terms used in ontology.
- Some of the methodologies are complex to build and takes long time and utilize large resources.
- The ontology building approach uses bottom-up approach which may not give exact components to merge to form ontology.
• With the reuse of ontology skeletal, there is no approach to reuse the existing ontology.

4.5.3 IMPLEMENTATION OF ONTOLOGY BUILDING FROM DATABASE

Usually the ontology building is performed manually, but researchers try to build ontology automatically or semi automatically to save the time and the efforts of building the ontology. We survey in this section the most important approaches that generate ontology from data.

There exists a group of definitions based on the procedure to build the ontology. These definitions highlight the relationships between ontology and knowledge bases. For example, the definition given by Bernaras and colleagues in the framework of KACTUS project: Ontology provides the means for describing explicitly the conceptualization behind the knowledge represented in a knowledge base. This definition proposes the extraction of ontology from a knowledge base (KB), which will give approach to build ontology. In this approach, the ontology is built following a bottom-up strategy on the basis of an application KB by means of an abstraction process [21].

Ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base. According to this definition, the same ontology can be used for building several KBs, which would share the same skeleton. These skeletons can be extended by adding low level sub-concepts or high level concepts that cover new areas. Such ontology gives easy and clear understanding of structure and inference mechanisms becomes easier [21].

4.5.3.1 Tools like Protégé Approach

There are certain tools are available like Protégé, Swoop, OntoEdit, Altova SemanticWorks, OntoStudio etc., which gives an environment to build ontology. It gives an ontology development with various aspects like superclass and subclass hierarchy, creating a sub class, instances for classes illustration, query retrieval process, graph representation etc. along with its attributes as concept, sub-concept and its relationships [26]. Superclass and subclass hierarchy are used to describe the concepts. We can add properties as slots and features as
facets to the classes to describe internal structure of these concepts. We can create slots by using editor and browser forms. Queries tab is used to run the query and find particular information about any instances or classes. It is a free, open-source platform that provides a suite of tools to construct domain models and knowledge-based applications with ontology. At its core, Protégé implements a rich set of knowledge-modeling structures and actions that support the creation, visualization, and manipulation of ontology in various representation formats. Protégé can be customized to provide domain-specific for creating knowledge models.

The Protégé platform supports two main ways of modeling ontology:

- The Protégé-Frames editor enables users to build and populate ontology that are frame-based, in accordance with the Open Knowledge Base Connectivity protocol (OKBC). In this model, ontology consists of a set of classes organized in a subsumption hierarchy to represent a domain's salient concepts, a set of slots associated to classes to describe their properties and relationships, and a set of instances of those classes - individual exemplars of the concepts that hold specific values for their properties [25].

- The Protégé-OWL editor enables users to build ontology for the Semantic Web, in particular in the W3C's Web Ontology Language (OWL). Given such ontology, the OWL formal semantics specifies how to derive its logical consequences, i.e. facts not literally present in the ontology, but entailed by the semantics. These entailments may be based on a single document or multiple distributed documents that have been combined using defined OWL mechanisms [25].

It does not use any fixed approach, but modification to an existing ontology is easier [22].

4.5.3.2 Data Mining Approach

Decision tree algorithm from data mining is used to discover and extract knowledge from structured data. Abd-Elraman Elsayed et al. [27] proposed a method for building ontology from the generated decision tree. Knowledge engineer uses ontology to represent the knowledge of domain expert. Decision tree gives the knowledge from the data in a prescribed
way. They map the data from decision tree as ontology terms such as concept, sub-concept and its relationships. Data has been transformed into extracted knowledge in the form of ontology by using data mining technique like classification. Finally, they mapped ontology with either XML and/or OWL format of data by using ontology builder module.

4.5.3.3 Ontology from RDBMS

There are many approaches proposed which defines semantics in database schema or extract semantics out of database schema. But these approaches does not exist hierarchies and cardinality about properties, so cannot be used to describe ontology. Irina et al. [28] proposed a methodology to store ontology in the form of SQL relational databases. The assumption has been made that the data is in third NF form. The process of acquiring ontology has been divided into two phases, first, extracting schema information and second, acquiring ontology. Each tuple of RDBMS has been tested for ontology instances for mapping. They designed some rules that mapped both [28].

4.5.3.4 Ontology to RDBMS

To store ontology there are many approaches proposed which are based on relational, object or object-relational. The relational are having advantages over object. Shufeng Zhou et al. [29] proposed the methodology of using transformation rules for ontology acquisition from relational database. Some rules have been proposed to map both. These rules depend on single-valued or multiple-valued that shows the number of ontology concepts mapped to number of tables or its entries. Finally, ontology has been represented in OWL form.
Table 4.2: Ontology Building Approaches from Database

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Name of Ontology Building Approach</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KACTUS Project</td>
<td>Bottom-up approach; ontology as a skeletal.</td>
</tr>
<tr>
<td>2</td>
<td>Protégé Tool</td>
<td>Editor to build ontology.</td>
</tr>
<tr>
<td>3</td>
<td>Data Mining Technique</td>
<td>Data from decision tree; XML OR OWL format.</td>
</tr>
<tr>
<td>4</td>
<td>SQL Relational DB</td>
<td>Rules to map.</td>
</tr>
<tr>
<td>5</td>
<td>Using Transformation Rules</td>
<td>Transformation rules that maps both.</td>
</tr>
</tbody>
</table>

4.5.4 **DRAWBACKS OF ONTOLOGY FROM DATABASE APPROACHES**

- Certain tools have been used to build ontology. Each one has its limitations or may not be suitable to build entire ontology for domain. The tool like protégé does not support synonyms.

- In decision tree approach, building of ontology will be cumbersome and time-consuming since it depends on training data.

- From literature survey, there are some rules that are not enough to transform relational database to ontology.

4.6 **DRAWBACKS OF EXISTING SEMANTIC MATCHING APPROACHES USING ONTOLOGY**

The following are some of the drawbacks from existing ontology for CLIR approaches.

- Most of the approaches use synsets from WordNet for each language. It may be time consuming to find out the synset from WordNet.

- Many approaches made semantic matching for documents instead of domain.

- Query expansion has been done to include all the terms of ontology to match from query terms. This may lead to time consumption to acquire information.
4.7 SUMMARY

In this chapter, the theoretical background needed for semantic matching has been explained. We elaborated the semantic matching approaches along with their matching strategies. A wide survey of some of the existing approaches for semantic matching using ontology has been made. Finally, ontology building approaches has been elaborated and found various drawbacks of those methodologies. Also, the drawback of semantic matching using ontology has been mentioned.