Implementation of Semantic Analysis Using Domain Ontology

Pratik Agrawal1, Prof. A.J.Agrawal [2]

[1, 2]*(Department of Computer Science, Shri Ramdeobaba College of Engineering & Management Nagpur, India)*

Abstract : *In feedback analysis of an organization, the organization wants to produce a summary of feedback based on organization entity. However for the same entity students can express it with different words and phrases. For a meaningful summary, these words and phrases which are domain specific needs to be grouped under the same entitygroup.This paper proposes an semantic based feedback analysis system that makes use of an semantic lexicon and organization ontology as a base for matching the specified entity with the help of a Jaccard similarity method. Experimental results using three different training dataset shows that the proposed method is competent for the task.*

Keywords *- Natural language processing, Ontology, Semantic lexicon, Semantic analysis*

I. INTRODUCTION

Semantics is the study of meaning of words. It concentrates on the study of relation between signifiers, like words, phrases and symbols, and what they stand for, their denotation. Semantic is essential for understanding language acquisition and change.

In social contexts English and effects of style are likely to change the meaning hence understanding language plays an important role**.** It is thus one of the most fundamental concepts in linguistics. Procedure of how meaning is constructed, interpreted, obscured, clarified, illustrated, simplified negotiated, contradicted and paraphrased is studied in semantics. Semantics tries to understand what meaning is as an element of language and how it is constructed by language as well as obscured, interpreted, and negotiated by speakers and listeners of language. Semantic information can be helpful in almost all aspects of natural language understanding, including word sense disambiguation, selectional restrictions, attachment decisions and discourse processing. Semantic knowledge can add a great deal of power and accuracy to natural language processing systems. But semantic information is difficult to obtain. Semantic processing determines the possible meanings of a sentence by focusing on the interactions among word-level meanings in the sentence. Processing of the semantic disambiguation of words with multiple senses is also included in semantic analysis. In an analogous way to how syntactic disambiguation of words that can function as multiple parts-of-speech is accomplished at the syntactic level. Semantic representation of the sentence consists of only one sense of selected words permitted by semantic disambiguation. For example, amongst other meanings, 'apple' as a noun can mean either as a fruit or it can be a company name.

Feedback is important for the organization and their evaluation plays an important role in the development. The feedback evaluation system consists of the semantic analysis for determining the entity of the organization. Semantic analysis is necessary for determining the words phrase with the organization entity and matching them with the help of Jaccard similarity based on the defined properties for the entity.

The paper is organizes as follows in the second part, we have talk about the motivation that we have got from the current problems for building up the feedback analysis system in the third part, we have described the work in semantic analysis that have been carried out in the fourth part, we have described various semantic analysis techniques in details in the fifth part, we have implemented the analysis system that is explained it's architecture steps and methods that have implemented for that in the sixth part, we have carried out the experimental evaluation on our system by taking three datasets and calculated Precision and recall finally in the seventh part, we have summarized and came to the conclusion.

II. MOTIVATION

The motivation of natural language based feedback system shall be described by an example. A person wants to provide the feedback for the organization. The organization will provide him with a feedback form. The feedback from consists of a radio button. The radio button is marked with some specified values based on some calculation. The person wants to click on the given radio button for providing feedback to the organization.

Due to the use of radio button, the person fails to express his feeling and also was not able to provide some suggestions regarding the organization. The organization is also not able to determine what the fault is in the system? And so judgment regarding about the problem is not made quickly.

So for all these problems, the feedback system is provided that will analysis person day to day words. So that the person can easily express his views in his own words and can provide valuable suggestions for the organization. The organization can easily detect the problems by the person suggestions and judgment regarding that problems can be made easily and quickly. This will helps to save time and will helps in better development of the organization.

III. RELATED WORK

Semantic analysis

The different works related to semantic analysis are specified below with the methods that are used for the implementation of the semantic analysis.

1. Mita K. Dalal and Mukesh A. Zaveri in their paper "Automatic Text Classification: A Technical Review" presents an Automatic Text Classification approach. Automatic Text Classification is a semi-supervised machine learning task that automatically assigns a given document to a set of pre-defined categories based on its textual content and extracted features. Automatic Text Classification involves assigning a text document to a set of pre-defined classes automatically, using a machine learning technique. The classification is usually done on the basis of significant words or features extracted from the text document. It is a supervised machine learning task so the classes aree already predefined. Most of the official communication and documentation maintained in commercial and governmental organizations is in the form of textual electronic documents and e-mails. Much of the personal and other communication done by private individuals is in the form of e-mails, blogs etc. Due to this information overload, efficient classification and retrieval of relevant content has gained significant importance. [9]

2. Ana-Maria Popescu and Oren Etzioni in their paper "Extracting Product Features and Opinions from Reviews" introduces OPINE.OPINE uses a novel relaxation-labeling technique to determine the semantic orientation of potential opinion words in the context of the extracted product features and specific review sentences; this technique allows the system to identify customer opinions and their polarity with high precision and Recall. Relaxation labeling is an iterative procedure whose output is an assignment of labels to objects. At each iteration, the algorithm uses an update equation to estimate the probability of an object label based on its previous probability estimate and the features of its neighborhood. The algorithm stops when the global label assignment stays constant over multiple consecutive iterations. [6]

3. Peter D. Turney in his paper "Thumbs Up or Thumbs Down? Of Semantic Orientation Applied to Unsupervised Classification of Reviews" introduces the PMI-IR algorithm. The PMI-IR algorithm is employed to estimate the semantic orientation of a phrase (Turney, 2001). PMI-IR uses Point wise Mutual Information (PMI) and Information Retrieval (IR) to measure the similarity of pairs of words or phrases. The semantic orientation of a given phrase is calculated by comparing its similarity to a positive reference word ("excellent") with its similarity to a negative reference word ("poor"). More specifically, a phrase is assigned a numerical rating by taking the mutual information between the given phrase and the word "excellent" and subtracting the mutual information between the given phrase and the word "poor". The disadvantage of the system is that it include the time required for queries and, for some applications, the level of accuracy that was achieved [7]

4. Vasileios Hatzivassiloglou and Kathleen R.Mckeown in this paper "Predicting the semantic orientation of adjectives" introduces the log-linear regression model. A log-linear regression model which combined with supplementary morphology rules predict whether conjoined adjectives are of the same or different orientations achieving 82% accuracy in this task when each conjunctions is considered independently. It combines information from different conjunctions to determine if each two conjoined adjectives are of the same or different types. [10]

5. Rada Mihalcea and Dan Moldovan in their paper "Semantic Indexing using WordNet Senses" proposed Boolean Information retrieval system that adds word semantics to the classic word based indexing. A combined word based and sense-based approach is used in indexing and retrieval components. The key to our system is a methodology for building semantic representations of free format text, at word and collocation level. This technique is called semantic indexing and shows improved effectiveness over the classic word based indexing techniques. The main problem with the traditional Boolean word-based approach to Information Retrieval (IR) is that it usually returns too many results or wrong results to be useful. Keywords have often multiple lexical functionalities (i.e. Can have various parts of speech) or have several semantic senses. Also, relevant information can be missed by not specifying the exact keywords. The solution is to include more information in the documents to be indexed, such as to enable a system to retrieve documents based on the words, regarded as lexical strings, or based on the semantic meaning of the words.

The Semantic indexing usage of word senses in the process of document indexing is a pretty much debated field of discussions. The basic idea is to index word meanings, rather than words take as lexical strings. Nevertheless, the conclusion which can be drawn from all these experiments is that a highly accurate Word Sense Disambiguation algorithm is needed in order to obtain an increase in the performance of IR systems. [8]

State of art:

Author	Model	Publication	Feature
Mita K. Dalal and Mukesh A. Zaveri	Automatic Text Classification	International Journal of Computer Applications (2011)	semi-supervised machine learning
Ana-Maria Popescu and Oren Etzioni	OPINE	Association for Computational Linguistics (2005)	novel relaxation-labeling technique
Peter D. Turney	PMI-IR algorithm	Computational Linguistics (ACL), Philadelphia (July 2002)	Point wise Mutual Information (PMI) and Information Retrieval (IR)
Vasileios Hatzivassiloglou andKathleen R.Mckeown	Log-linear regression model	Published in (1997)	morphology rules
RadaMihalcea andDanMoldovan	Semantic indexing	Published in (2000)	classic word based indexing

IV. TECHNIQUES AND METHODS

There are various semantics method available today we will described some semantic analysis types in detail so that it can be helpful for the analysis.

Corpus based approach:

Corpus-based approaches use automated learning techniques over corpora of natural language examples in an attempt to automatically induce suitable language-processing models. Corpus-based semantic models adopt the general idea that similar words will be used in similar contexts and represent the meaning of a word/concept as a vector (ordered list) of values that summarize, the context signature of the word, in some way that is, the set of contexts in which the word appears. In a high-dimensional space, words can be represented as points with the help of these vectors, and the distance between words is measured using standard geometric techniques, such as computing the cosine of the angle between two vectors.

Traditional work in natural language systems breaks the process of understanding into broad areas of syntactic processing, semantic interpretation, and discourse pragmatics. Most empirical NLP work to date has focused on using statistical or other learning techniques to automate relatively low-level language processing such as part-of-speech tagging, segmenting text, and syntactic parsing. The success of these approaches, following on the heels of the success of similar techniques in speech-recognition research, has stimulated research in using empirical learning techniques in other facts of NLP, including semantic analysis that is used for uncovering the meaning of an utterance.[14]

WordNet approach:

WordNet is a semantic network, which is organized in such a way that synset and word senses are the nodes of the network, and relations among the synset and word senses are the edges of the network. In WordNet, each meaning of a word is represented by a unique word sense of the word, and a synset (stands for “synonym set”) is consisting of a group of word senses sharing the same meaning. More than two thirds of the nodes in WordNet are synset. Hyponym Of is the key relationship for noun synset in WordNet, which has been widely used to estimate the semantic relatedness among nouns.

WordNet has been commonly used to measure semantic similarity among words since it has the inherent advantages of being structured in the way of simulating human recognition behaviors. [13]

Ontology:

Ontology takes the role as a philosophical concept originally, which regarded as an explanation or description to an objective existing system from the philosophy category, and concerns the abstract nature of the objective reality. Later, it is given a new definition by the artificial intelligence category with the development of artificial intelligence.

The Artificial-Intelligence literature contains many definitions of ontology; many of these are not similar to each other. For guiding this purpose an ontology is a formal explicit description of concepts in a domain of discourse (classes), properties of each concept describing various features and attributes of the concept (slots), and restrictions on slots (facets (sometimes called role restrictions)).Ontology in combination with a set of individual instances of classes constitutes a knowledge base. In actual practical, there is a fine line where the ontology ends and the knowledge base begins.

In practical terms, developing Ontology includes:

- Defining classes in the ontology,
- Arranging the classes in a taxonomic (subclass–super class) hierarchy,
- Defining slots and describing allowed values for these slots,
- For instances values in the slots are filled.[15]

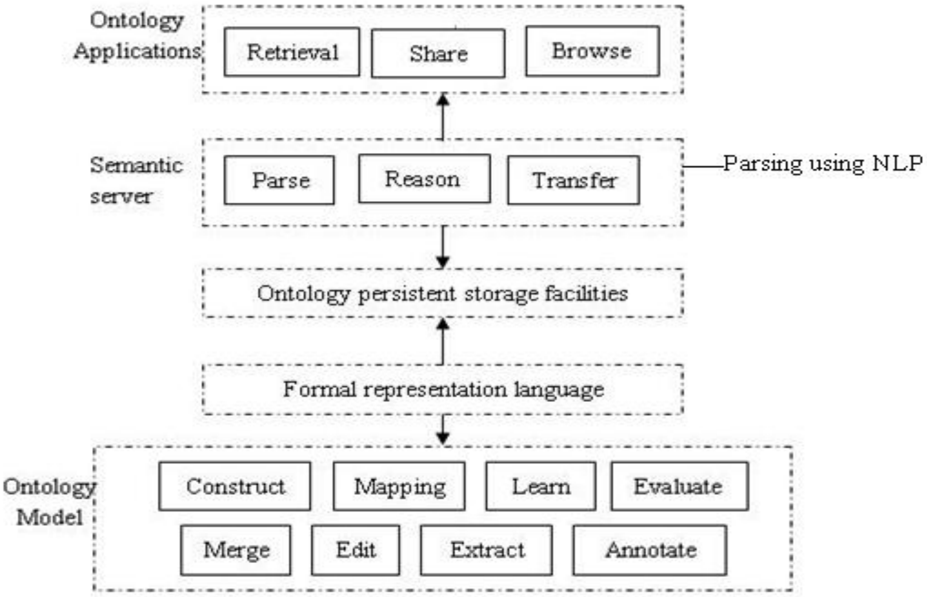

Fig: 1.Architecture of ontology-based applications

V. PROPOSED WORK

Architecture of feedback analysis system

The design and implemented architecture of the feedback analysis system are as follows. The system analyzes all the input feedback and provide with all the semantic matching analysis with the entity ontology. The architecture is explained with each and every process that is going to take place**.**

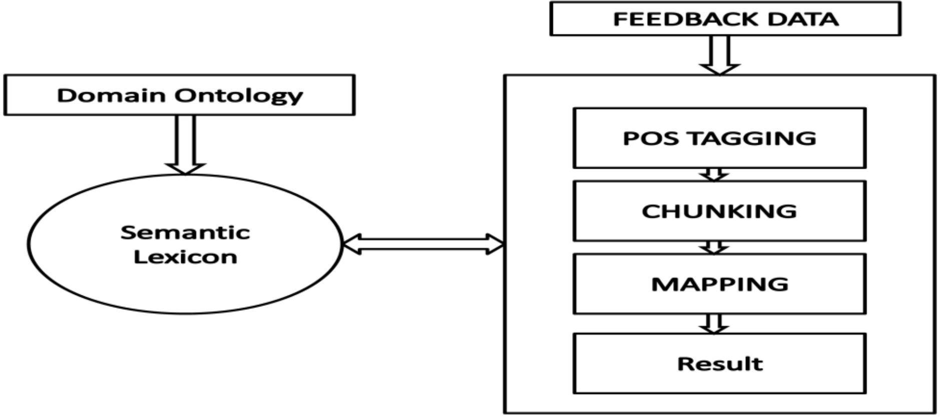

Fig: 2 Architecture of feedback analysis system

5.1 Data collection

The college is considered as our organization. The feedback forms are provided to all the student of various departments and have collected the feedback. Like this the set of input data is created. The data collected was in an unstructured from we have taken that data and extracted the information data that is needed for our work. There were various guidelines for feeling the form the guidelines were given to them base on the organization. By providing the guidelines it was easy for the students to provide the proper feedback. The guidelines given to them were college, canteen, library, lab facility, extracurricular activities, teaching. The students have considered these guidelines as a base and have provided the feedback for the college.

The input data (feedback) collected from the students will help in analyzing and generating the result for the organization.

5.2 Preprocessing of the feedback

Part-of-speech tagging is often a critical first step in various speech and language processing tasks. High-accuracy taggers (e.g., based on conditional random fields) rely on well chosen feature functions to ensure that important characteristics of the empirical training distribution are reflected in the trained model. This makes them vulnerable to any discrepancy between training and tagging corpora, and accuracy is adversely affected by the presence of out-of-vocabulary words.

In corpus linguistics, part-of-speech tagging also called grammatical tagging or word-category disambiguation, is the process of marking up the words in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context —i.e. relationship with adjacent and related words in a sentence, phrase, or paragraph. Part-of-speech tagging is harder than just having a list of words and their parts of speech, because some words represent more than one part of speech at many times, and because some parts of speech are complex or unspoken. So the pos tagging is important in the semantic analysis there are various types of tagger available for tagging. The various taggers like Stanford tagger, Latent Analogy, POSLDA, OPINE and based on their accuracy on Penn Treebank bank analysis the Stanford tagger was selected that give an accuracy of 97.24%. So we have implemented this pos tagger in our feedback analysis system.

The feedbacks collected from the students are analyzed by the method of Stanford tagger. A Part-Of-Speech Tagger (POS Tagger) that reads text in some language and assigns parts of speech to each word such as noun, verb, adjective, etc

The Stanford tagger is trained with 10, 00,000 words from the Oxford dictionary with the help of some standard pre-defined methods. The tagger uses the following ideas (i) explicit use of both preceding and following tag contexts via a dependency network representation, (ii) it uses broad view of lexical features, including jointly conditioning on multiple consecutive words, (iii) effective use of priors in conditional log linear models, and (iv) fine-grained modeling of unknown word features. Using these ideas together, a bi-directional dependency network tagger in bidirectional/wsj3t0-18.holder gives 97.24% accuracy on the Penn Treebank WSJ, an error reduction of 4.4% on the best previous single automatically learned tagging result.The tagger uses a bi-directional dependency network tagger for tagging the words the tagger is composed of both the features of tagging it uses a CMM method of left to right and right to left for extracting the tagged tokens. We proposed a Maxnet tagger method that makes use of a bi-directional dependency network tagger and tokenize the sentence and assigns a part of speech to each word of a sentence. The tagging is important to distinguish the noun phrase from the sentence as well as adjectives words that will help in analyzing the semantic analysis as well as sentiment analysis of the feedback. [1]

For example

Sports facilities are not available.

Pos tagged

Sports/NNS facility/NN are/VBP not/RB available/JJ.

5.3 Organization Ontology models

Researchers have shown that it's hard to describe college feedback concepts and their relationships using only one college ontology model, it's possible to construct sub domain ontology models in organization sector. The mostly used ontology construction methods including Skeleton method, TOVE ontology, G&FOX method, KACTUS And Bernaras methods, SENSUS, IDEF5 and seven-step method. Ontolingua, Ontodaurus, WebOnto, Protégé, OntoEdit are the common tools to construct ontology models. We have practiced constructing college chain ontology model used seven-step method.

The seven-steps are:

1. Determine areas scope of domain ontology;
2. Examine the possibility of using existing ontology;
3. List important terms in the ontology;
4. Define classes and class hierarchy;

5. Define class attributes;
6. The definition of property distribution;
7. Create instances;

The organization entity are design in the form of ontology it is necessary so that we can easily find out the entity relationship with the other entity. The designing helps us to known easily which entity belongs to which one of the entity and their property helps us to map the specified entity. We have design the domain ontology of our organization that will help us in determining the specified entity based on their property.

Representation of ontology models

Ontology formal representations are explicit description of Ontologies formulated in terms of a given ontology (description) language. There have some ontology formal languages such as RDF (Resource Description Framework), RDF(S), OIL, DAML, OWL (DAML+OIL is regarded as a Transition), KIF, SHOE, XOL, OCML, Ontolingua, Cycle and Loom.

College ontology model

The college ontology is the basic and foremost ontology that we have described it comprises of all the entity that are related to the college and cover almost all of the small and big entities within it. The entities are distinguished based on their properties to which they contain and are related. The college ontology contains the infrastructure, canteen, library, lab facility, extracurricular activities, teaching, and placement. These all are the basic entity that the college is having and based on these entity feedbacks are classified and are related to that entity if they match their specific properties. We have created sub domain ontology of each entity so that we can get a detail and brief introduction about their properties.

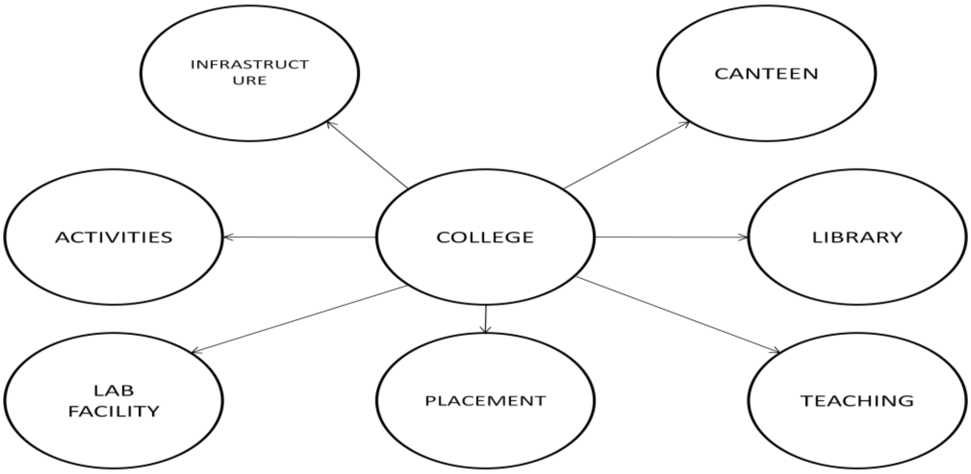

Fig: 3 College ontology model

Canteen ontology model

The canteen ontology is sub domain ontology it express the canteen entity in detail according to their desires properties. The canteen entity consist of the properties like food quality, infrastructure, hygienic, staff these all the properties define the canteen entity completely. The each property contain various attributes that are described in the lexicon so that it will be helpful for the classification and matching of the words or phrases extracted in the pre processing stage and is helpful in classifying the entity.

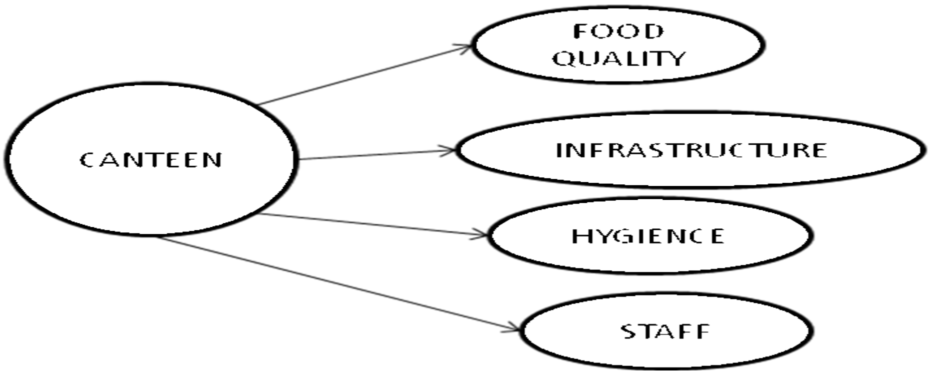

Fig: 4. Canteen ontology model

Lab facility ontology model

The lab ontology is sub domain ontology it express the lab entity in detail according to their desires properties. The lab facility consists of the properties like timing, staff, equipment, internet staff these all the properties define the lab entity completely. The each property contain various attributes that are described in the lexicon so that it will be helpful for the classification and matching of the words or phrases extracted in the pre processing stage and is helpful in classifying the entity

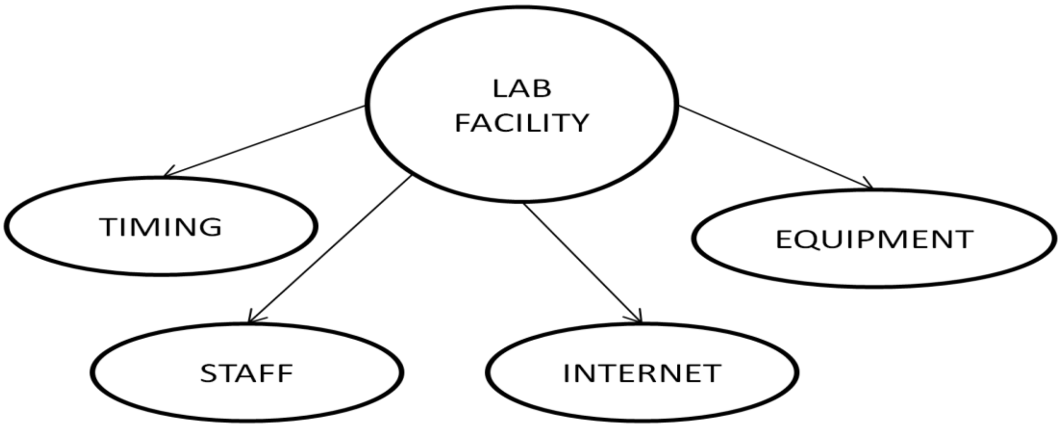

Fig: 5 Lab facility ontology model

Placement ontology model

The placement ontology is sub domain ontology it express the placement entity in detail according to their desires properties. The placement entity consist of the properties like package, training, industry exposure these all the properties define the placement entity completely. The each property contain various attributes that are described in the lexicon so that it will be helpful for the classification and matching of the words or phrases extracted in the pre processing stage and is helpful in classifying the entity.

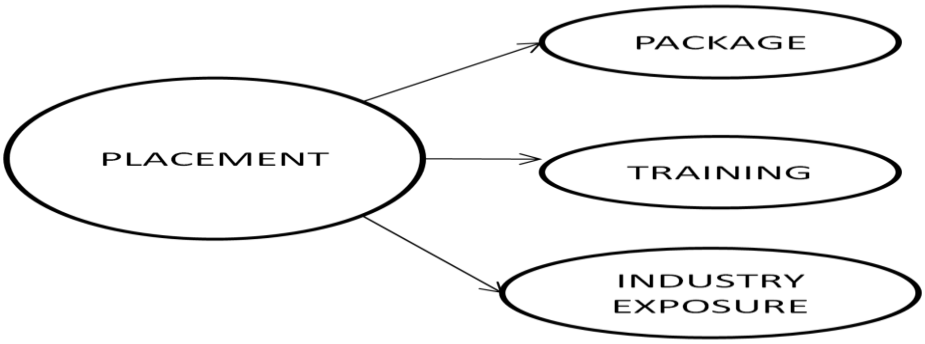

Fig: 6 Placement ontology model

Library ontology model

The library ontology is sub domain ontology it express the library entity in detail according to their desires properties. The placement entity consist of the properties like books, staff, timings, infrastructure these all the properties define the library entity completely. The each property contain various attributes that are described in the lexicon so that it will be helpful for the classification and matching of the words or phrases extracted in the pre processing stage and is helpful in classifying the entity

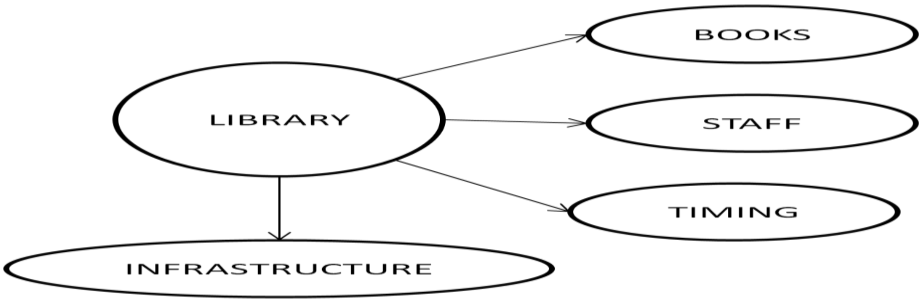

Fig: 7 library ontology model

Teaching ontology model

The teaching ontology is sub domain ontology it express the teaching entity in detail according to their desires properties. The teaching entity consist of the properties like timings, evaluation, contents, faculty these all the properties define the teaching entity completely. The each property contain various attributes that are described in the lexicon so that it will be helpful for the classification and matching of the words or phrases extracted in the pre processing stage and is helpful in classifying the entity.

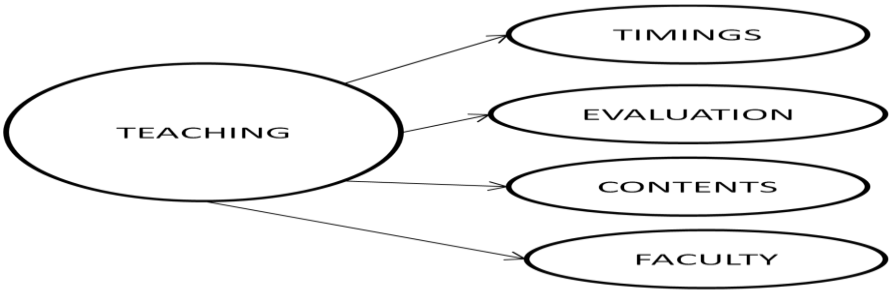

Fig: 8 teaching ontology model

Infrastructure ontology model

The Infrastructure ontology is sub domain ontology it express the Infrastructure entity in detail according to their desires properties. The Infrastructure entity consist of the properties like classroom, security, parking, washroom, space these all the properties define the Infrastructure entity completely. The each property contain various attributes that are described in the lexicon so that it will be helpful for the classification and matching of the words or phrases extracted in the pre processing stage and is helpful in classifying the entity.

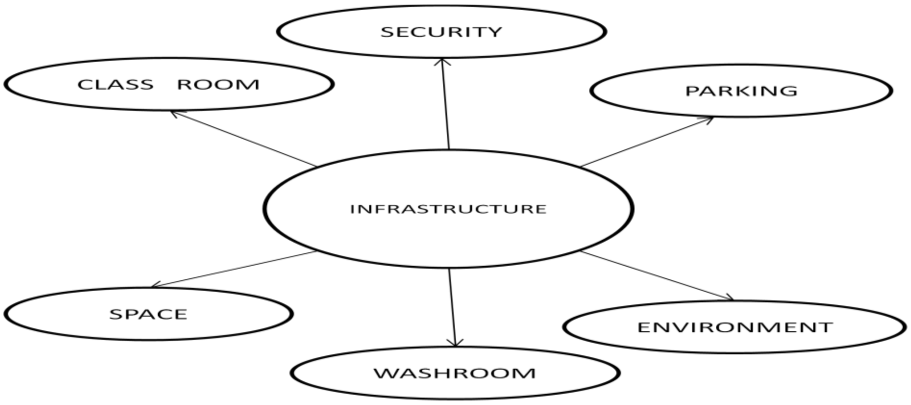

Fig: 9 Infrastructure ontology model

Extracurricular ontology model

The extracurricular ontology is sub domain ontology it express the extracurricular entity in detail according to their desires properties. The extracurricular entity consist of the properties like encouragement, trainer, accessories, financial these all the properties define the extracurricular entity completely. The each property contain various attributes that are described in the lexicon so that it will be helpful for the classification and matching of the words or phrases extracted in the pre processing stage and is helpful in classifying the entity.

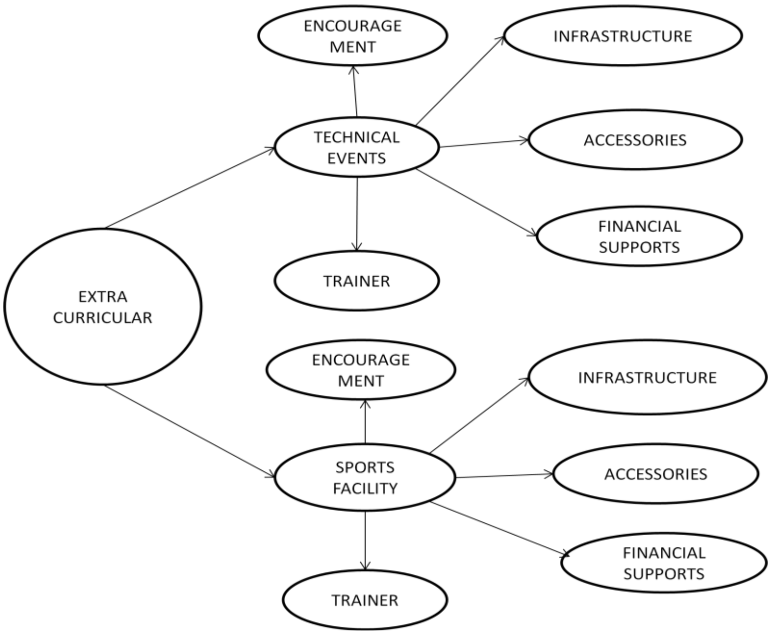

Fig: 10Extracurricular ontology model

5.4 Chunking of feedback data.

Syntax captures structural relationships between words and a phrase i.e., describes the constituent structure of NL expressions. Grammars are used to describe the syntax of a language Syntactic analysers assign a syntactic structure to a string on the basis of a grammar. A syntactic analyser is also called a parser.

Chunking is partial parsing. A chunker assigns a partial syntactic structure to a sentence. It yields flatter structures than full parsing. The chunking only deals with “chunks” takes only the tagged words that are defined by the grammar.

The chunking uses deterministic grammars for easy-to-parse pieces, and other methods for other pieces, depending on task. Chunking is basically the identification of parts of speech and short phrases (like noun phrases). Part of speech tagging tells you whether words are nouns, verbs, adjectives, etc.The chunking takes place by combining all the terms that are specified in the grammar. We have proposed a regular expression grammar that defines the NP words from the sentences. We can call it a bag of words. These NP words or phrases will be used to match the words from the semantic lexicon and in order to determine which NP is related to which entity of the organization based on the properties.

From the pos tagged words we have analyzed and extracted all the NP words with the help of grammar that defines the entity of the organization.

For example:

Sports/NNS facility/NN are/VBP not/RB available/JJ. /.

Chunked NP word

(NP Sports/NNS facility/NN)

5.5 Semantic lexicon

A lexicon is a list of words in a language—a vocabulary—along with some knowledge of how each word is used. A lexicon may be general or domain-specific; we might have, for example, a lexicon of several thousand common words of English or some language. The words that are of interest are usually open-class or content words, such as nouns, verbs, and adjectives, rather than closed-class or grammatical function words, such as articles, pronouns, and prepositions, whose behavior is more tightly bound to the grammar of the language. A lexicon may also include multi-word expressions such as fixed phrases (by and large), phrasal verbs (tear apart), and other common expressions.

Each word or phrase in a lexicon is described in a lexical entry; exactly what is included in each entry depends on the purpose of the particular lexicon. The details that are given may include any of its properties of spelling or sound, grammatical behaviour, meaning, or use, and the nature of its relationships with other words. A lexical entry is therefore a potentially large record specifying many aspects of the linguistic behaviour and meaning of a word.

A lexicon can -be viewed as an index that maps from the written form of a word to information about that word. This is not a one-to-one correspondence, however. Words that occur in more than one syntactic category will usually have a separate entry for each category; for example, flap would have one entry as a noun and another as a verb. Separate entries are usually also appropriate for each of the senses of a homonym-refers to lexemes with the same form but unrelated meanings and the second term polysemy –refers to the notion of a single lexeme with multiple related meanings. A lexicon may be just a simple list of entries, or a more-complex structure may be imposed upon it. For example, a lexicon may be organized hierarchically, with default inheritance of linguistic properties.

Lexical entries include the linguistic behaviour or use of a word its phonetics, morphology, written forms, behaviour; it's relative frequency and all other aspects of its meaning. The word semantic properties include relationship between the meaning of the word and those of other words. The lexicon possesses the inheritance properties we can inherit a lexicon by other one. The "classical" lexical relationships pertain to identity of meaning, inclusion of meaning, part–whole relationships, and opposite meanings. Identity of meaning is synonymy: Two or more words are synonyms (with respect to one sense of each) if one may substitute for another in a text without changing the meaning of the text.

The lexicon has a highly semantic structure that governs what words can mean, and how they can be used. This structure consists of relations among words and their meanings, as well as the internal structure of individual words. The linguistic study of this systematic, meaning related, structure is called lexical semantics. We have used the lexeme, an individual entry in the lexicon. A lexeme should be thought of as a pairing of a particular orthographic and phonological form with some form of symbolic meaning representation. The lexicon is a finite set of lexemes. This allows us to include compound nouns and other compositional phrases as entries in the lexicon.

We have proposed and created a semantic lexicon based on the feedback data. The semantic lexicon includes the organization entities their properties as well as the words or phrases that define the entity with their properties. The words or phrases are extracted from the pos tagged data by forming the chunk of the data with the help of predefined grammar. The format for the database file is words, entity, properties.

5.6 Similarity matching

Measures of semantic similarity between concepts are widely used in Natural Language Processing. Similarity evaluation between two documents is an important operation which lies at the heart of most text and language processing tasks. The similarity evaluation forms a main part of the information retrieval system for retrieving the information. We need some parameters to be similar so that we can retrieve the similar thing from the large bunch of materials. The similarity matching are used at many places like, when we make use of a search engine, we request web-page documents which bear some similarity to the keywords or string(s) which constitute the query document, when we ask for a text in some language A to be translated into some language B, we request a document in language B which has some similarity to the document in language A, when we summaries one document, we seek to produce another document which is similar in some way, which however is also different in other ways. Naturally, such use of document similarity could easily involve other media than text, such as pictures, audio, etc. However, the metrics for such media can understandably be very different (and often more primitive) from the ones we will be considering in text processing.

Considering the above example by the use of similarity matching we are able to compute the partial goal of the system like classification, validation, generation, etc. A measure of semantic similarity takes as input two words or phrases, and returns a numeric score that quantifies how much they are alike. Such a measure is usually based on 'is –a' relations found in the underlying taxonomy or ontology in which the words or phrases reside. For example, drinking water and water are similar in that a drinking water is a kind of water. Likewise, drinking water and water cooler are similar in that they are both kinds of water. Of course, many ontologies include additional relations between concepts such as has-part, is-a-way-of-doing, is-belongs-to etc that are not directly accounted for in measures of similarity.

Thus, we view semantic similarity as a special case of semantic relatedness, and we believe that developing measures that take advantage of increasingly rich ontologies (particularly in the organization domain) which have a wealth of relations beyond is an important area of research.

There are different methods available for similarity matching. In our system we are not able to match the exact similarity between the pos tagged words and the semantic lexicon because there are various human spelling error that have to consider while evaluating similarity. Avoiding the human spelling error problem can become a great bottle neck for our system because it is not possible that all students will spell the words properly as described in the lexicon and by considering exact similarity match it will not be possible for the system to match that word to the lexicon and all that words will not be considered resulting in degrading the performance of the feedback evaluation system.

So to solve this problem in the system implementing some similarity method that matches the words and provide some similarity coefficient value so that by setting up some threshold value it will be possible to consider all the matching entries for the further evaluation and like this the problem of human spelling error can be overcome. There are various similarity evaluation methods such as cosine similarity, Jaccard similarity. In our system we had implemented Jaccard similarity.

Jaccard similarity determines the Jaccard coefficient. The Jaccard similarity coefficient is a statistical measure of similarity between sample sets and it is defined as the cardinality of their intersection divided by the cardinality of their union.

Mathematically,

$J (A, B) = |A \cap B| / | A \cup B|$

Eg: $X= \{A, B, C, D, E\}, Y = \{B, C, D, E, F\}$

X and Y are words.

Jaccard similarity=4/6=0.67

In our system first we have extracted the pos tagged words from all the feedback sentences from that we have chunk the NP words from all the sentences based on the chunking grammar that was defined for our system. The similarity match between the chunked words and the semantic lexicon is then calculated. The first column in the semantic lexicon contains the predefined words for the entity. We have taken the chunked word and had matched that word with the first column of the lexicon and for each match we have got the Jaccard coefficient value. Word pair with similarity score more than 0.7 are considered as similar. Second column of semantic lexicon describes entity and entity corresponding to the similar word is used further processing.

For example Chunked words are Drinking water, Food quality, sequrity, teaching for these words, and lexicon first column words are water, food quality, security, teachers which have Jaccard coefficient more than 0.7 and corresponding entities are canteen, canteen, infrastructure, and teaching. We have matched that word to the specified entity from the semantic lexicon so the final result is in terms of related to part. Drinking water is related to canteen, food quality related to canteen, security related to infrastructure and teaching faculty related to teaching. Like this all the NP chunked words are matched to their entity of the organization with the help of Jaccard similarity method.

The result that we have got by the similarity matching are compared based on the different training set of the semantic lexicon.

VI. EXPERIMENTAL RESULT

Since the quality of many retrieval systems depends on how well they manage to rank relevant documents before non-relevant ones, IR researchers have developed evaluation measures specifically designed to evaluate rankings. Most of these measures combine precision and recall in a way that takes account of the ranking.The result of the semantic analysis are computed by taking the parameters precision and recall.

Precision is one measure of the effectiveness of some computer applications for finding search words, candidate terms, and other items. Precision is a measure of the proportion of results of a computer application that are considered to be pertinent or correct. For example, if an application is searching for terms in a document and finds 100 candidates, 75 of which really are terms (that is, there are 75 correct results out of 100 total), then the application's Precision is 0.75.

Recall is one measure of the effectiveness of some computer applications for finding search words, candidate terms, and other items. Recall is a measure of the proportion of all possible correct results of a computer application that the application actually produces. For example, imagine that you are using a computer application to search for terms in a document that has 80 terms in it. (You know because you counted them.) If the application finds 55 of these terms, then the recall of the application is 55 out of 80, or 0.62.

Obviously, recall is not a measure that is applied every time you use a computer application, because it would defeat the purpose to count all the correct results yourself before using a tool! However, testing recall can help to evaluate a tool's performance before you implement it and determine whether the tool can be useful in other jobs and will help to adjust settings on different tools to adjust their performance to meet the specified needs. [12]

In feedback system the result of precision and recall are calculated by training the semantic lexicon by the different data set. For the experiment the three dataset are considered and have trained each time our lexicon with different dataset and have test that on test dataset and have calculated the result of precision and recall.

In the first data set the semantic lexicon was trained with 500 sentences and has created a semantic lexicon based on the entity and the phrases mentioned in the training sentences. All the words from the sentences are extracted that emphasis the entity of the organization based on their properties. The words are related to their entity by some specific properties they satisfies each entity is having some predefined properties so it is easy for the similarity matching. The similarity match is carried out by using the Jaccard similarity match that uses the defined formula and calculate the similarity match between the extracted words and the semantic lexicon and give the desired result. A threshold of 0.6 is set so it will be easy for the maximum entries to be matched with the semantic lexicon.

In the second dataset, semantic lexicon was trained by 800 sentences means more entity words and their terms are added in that lexicon in the first data set lexicon was not able to map some words to their specific entity so by training it with more 300 hundred sentences it will be able to map more words to semantic lexicon gradually increasing our result.

In the third data set the semantic lexicon was trained with 1200 sentences more words phrases and their related entity in the lexicon are added making our semantic lexicon more precise and able to handle all the words coming from the test dataset improving the above result therefore by training with additional 400 sentences result in the more accuracy.

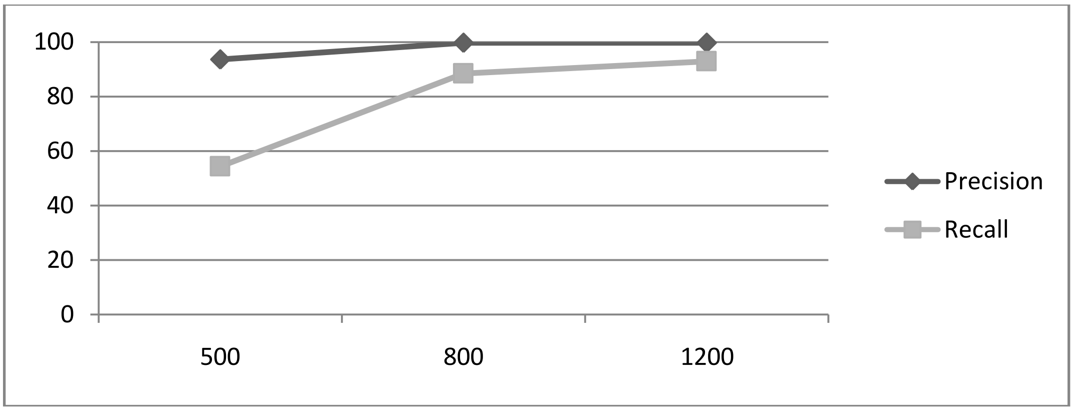

Fig: 11 Precision and recall

SR NO	TRAINING DATASET	PRECISION	RECALL
1	500	93.65	54.29
2	800	99.65	88.34
3	1200	99.67	92.94

Finally the result of the Precision and recall for our first training dataset, a related match was 177 and the unrelated match was 12 and total unmatched in the database was 137 by taking this under consideration the value of Precision is 93% and of recall is 54%.

Likewise the Precision and recall for our second training dataset are calculated by training the dataset with more 300 sentences and got a considerably changed in the result the related match of 288 and the unrelated match of 1 and the total unmatched in the database of 37 by taking this into consideration the value of Precision is 99% and of recall is 88%.For the third training dataset the result were quite improved a related match of 303 and the unrelated match of 1 and the total unmatched in the database is 22 and calculated the Precision of 99% and the recall of 92% and at last we have plot the result on the graph.

VII. CONCLUSIONS

In this paper the semantic analysis its type and how it is useful in our feedback analysis system are explained and the steps that are required to carry out the semantic analysis as well as the proposed architecture of the system. We have also explained the preprocessing, data collection, ontology designed, chunking, semantic lexicon steps that we have carried out and at last we have implemented the similarity match by using the Jaccard similarity method and have calculated the result based on the three training dataset of 500,800,1200. We have calculated the Precision and recall base on this training dataset and have plotted the result on the graph for better understanding. We have come to known that the more you trained your semantic lexicon the better the similarity match will be possible.

REFERENCES

[1] Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network Kristina Toutanova ,Dan Klein Computer Science Stanford University Stanford, CA 94305-9040 kristina@cs.stanford.edu ,klein@cs.stanford.edu Manning (2011)

[2] Automatic Text Classification: A Technical Review Mita K. Dalal and Mukesh A. Zaveri International Journal of Computer Applications (0975 – 8887) Volume 28– No.2, August 2011.

[3] Machine Learning for Sentiment Analysis on the Experience Project Raymond Hsu Computer Science Dept. Stanford University hsuray@cs.stanford.eduBozhiSeeElectricalEngineeringDept.StanfordUniversitybozhi@stanford.eduAlanwuElectricalEngineeringDept.StanfordUniversityalanw@stanford.edu published in 2011.

[4] Survey on Feedback Analysis using Domain Ontology. International Conference on Computer Science & Information Technology, 20th Jan-2013, Mumbai, ISBN: 978-93-82208-56-3. Pratik agrawal Department of Computer Science, Shri Ramdeobaba College of Engineering & ManagementNagpur, India Prof. A.J.Agrawal Shri Ramdeobaba College of Engineering & ManagementNagpur, India

[5] Ontology-based Natural Language Processing for In-store Shopping Situations Sabine Janzen Furtwangen University Furtwangen, Maass FurtwangenUniversityFurtwangen,Germanywolfgang.maass@hs-furtwangen.de

[6] Extracting Product Features and Opinions from Reviews Ana-Maria Popescu and Oren Etzioni Department of Computer Science and Engineering University of Washington Seattle, WA 98195-2350 {amp, etzioni}@cs.washington.edu Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 339–346, Vancouver, October 2005. c 2005 Association for Computational Linguistics.

[7] Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews Peter D. Turney Institute for Information Technology National Research Council of Canada Ottawa, Ontario, Canada, K1A 0R6 peter.turney@nrc.ca Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424

[8] Semantic Indexing using WordNet Senses Rada Mihalcea and Dan Moldovan department of computer science and engineering southern Methodist university dallas texas 75275-0122 in 2000

[9] Automatic Text Classification: A Technical Review Mita K. Dalal and Mukesh A. Zaveri International Journal of Computer Applications (0975 – 8887) Volume 28– No.2, August 2011.

[10] Predicting the semantic orientation of adjectives. Vasileios Hatzivassiloglou and Kathleen R.Mckeown published in 1997.

[11] Foundations of Statistical Natural Language Processing by Christopher D. Manning, Hinrich schutze

[12] Measures of semantic similarity and relatedness in the biomedical domain Ted Pedersen a,*, Serguei V.S. Pakhomov b, Siddharth Patwardhan c, Christopher G. Chute b a Department of Computer Science, 1114 Kirby Drive, University of Minnesota, Duluth, MN 55812, USAb Division of Biomedical Informatics, Mayo College of Medicine, Rochester, MN, USA c School of Computing, University of Utah, Salt Lake City, UT, USA

[13] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press,1998.

[14] Corpus-Based Approaches to Semantic Interpretation in Natural Language Processing AI Magazine By Ng, Hwee Tou; Zelle, John Magazine article from AI Magazine, Vol. 18, No. 4

[15] Ontology Based DistributedAgricultural Knowledge Management Xiuqin Qiu College of computer science & technology, Huazhong University of science & technology, WUHAN, CHINA Jun Yue*College of information science and technology,Ludong University, YANTAI, CHINA