DOI: 10.21917/ijsc.2015.0126

WORD SENSE DISAMBIGUATION METHOD USING SEMANTIC SIMILARITY MEASURES AND OWA OPERATOR

Kanika Mittal[1] and Amita Jain[2]

[1]*Department of Computer Science and Engineering, Bhagwan Parshuram Institute of Technology, India*
E-mail: kkanika_virgo@yahoo.com

[2]*Department of Computer Science, Ambedkar Institute of Advanced Communication Technologies & Research, India*
E-mail: amitajain@aiactr.ac.in

Abstract

Query expansion (QE) is the process of reformulating a query to improve retrieval performance. Most of the times user's query contains ambiguous terms which adds relevant as well as irrelevant terms to the query after applying current query expansion methods. This results in to low precision. The existing query expansion techniques do not consider the context of the ambiguous words present in the user's query. This paper presents a method for resolving the correct sense of the ambiguous terms present in the query by determining the similarity of the ambiguous term with the other terms in the query and then assigning weights to the similarity. The weights to the similarity measures of the terms are assigned on the basis of decreasing order of distance to the ambiguous term. An aggregated similarity score based on the assigned weights is calculated using Ordered weighted averaging operator (OWA) for each sense of the ambiguous term and the sense having highest similarity score will be considered as the most appropriate sense for the ambiguous term. After then, the query is expanded by taking an implicit feedback from user and adding terms related to the corresponding sense and hence optimizing the query expansion process.

Keywords:

Query Expansion, Natural Language Processing, Information Retrieval, Word Sense Disambiguation, WordNet, OWA Operator

1. INTRODUCTION

Information retrieval [1] is a process of retrieving the relevant documents from the document database when the user enters his query in the search engine. But sometimes, it results in retrieval of irrelevant documents also along-with relevant documents as the user is unclear about the information actually needed. The uncertainty in the user’s query induces ambiguity due to which inappropriate documents are retrieved. Also heterogeneous and dynamically changing information are major challenges of web data [2] which results in low precision. In order improve the retrieval efficiency, Query expansion technique is used in which user’s query is modified by addition of certain terms into the original query specific to the context of the query terms i.e. the sense of the terms present in the query as single word can have more than one meaning. Through query expansion, the ambiguity of terms can be dealt and the effects of the word mismatch problem are reduced which is a result of different terms being used in reference to a single concept, both in the documents and in the user queries. The other challenge in the fields of Natural Language Processing (NLP) and Information Retrieval (IR) is characteristics of polysemy and synonymy that exist in words of natural language. Human language is ambiguous, so the words can be interpreted in multiple ways based on the context in which they occur i.e. there can be more than one meaning of a particular word. For example:-

He went to *bank* to withdraw cash.

He went to *bank* to catch fishes.

The occurrences of the word bank in the two sentences clearly denote different meanings: A financial bank and a type of river bank, respectively.

In many cases, even human being naturally intelligent found it difficult to determine the intended meaning of an ambiguous word, they do not think about the ambiguity of the language and so it is extremely difficult to replicate this process computationally. So in this regard there is a need for a consistent computational model to assess this type of relation. When a word level semantic relation requires exploration, there are many potential types of relations that can be considered: hierarchical (e.g. IS-A or hypernym-hyponym, part-whole, etc.), associative (e.g. cause-effect), equivalence (synonymy), etc. [4]. For many cases, it has been suggested to study a special case of semantic relations- semantic similarity between the words or semantic distance [4]. A number of measures of semantic relatedness have been developed and evaluated by different researchers, which can be used to evaluate semantic similarity between nodes in the network. There are broadly two methods for measuring word similarity, in terms of utilizing knowledge base, which are classified into knowledge-rich and knowledge-poor methods [5]. The knowledge-rich approaches require either a tagging of words or conceptual dependency representation. Most methods that calculate semantic distance through WordNet [6] fall into this category; while in the knowledge poor approaches there is no need of encoded semantic information, and it solely depend on frequency of co-occurrence of word contexts to determine similarity. Most of the semantic relatedness measures which have been proposed in recent years rely on the noun taxonomy of the lexical database WordNet and can be classified into three categories namely [7]:

1) Edge-counting methods: The edge-counting or shortest path method according to quilling semantic memory model [8] finds the number of hops between the nodes determines the similarity of the concepts. Leacock and Chodorow [6] proposed the concept of information content [9] in to evaluate the relatedness of two words using the following model: $Sim\ (W_i,\ W_j) = [-log\ Dist\ (c_i,\ c_j)\ /\ 2 \times D]$, where, $Dist\ (c_i,\ c_j)$, is the shortest distance between concepts c_i and c_j; D is the maximum depth of the taxonomy. According to the claims of Wu and Palmer [10], the relatedness of two words is related to each other by the weighted sum of all their senses comparison.

2) Corpus based methods: According to Resnik [9] each synset in WordNet is associated with an information content value derived from corpus of text. The relatedness between two concepts is measured by the information content value of the most specific concept that the two concepts have in common.
3) Gloss based measures: Lesk [5] defines relatedness between two terms by taking into account, the dictionary definition overlaps of concepts. He describes an algorithm that disambiguates words based on the extent of overlaps of their dictionary definitions with those of words in the context.

In order to determine which sense of a word is activated by its use in a particular context; Word Sense Disambiguation (WSD) is used [3]. We can describe WSD as the task of assigning the appropriate sense(s) to all or some of the words in T. This is basically a method which make use of dictionary definitions (or glosses) of surrounding words to determine the correct sense of a particular word. The word which needs to be disambiguated is designated as the target word, and the surrounding words (to the target word) in the text as the context. One of the algorithm based on this approach known as Lesk Algorithm [5] which determines the sense of the target word whose dictionary definition has the maximum overlap of words with definitions of senses of other words in the context is the sense that is selected as the intended sense of the target word.

In literature, several approaches have been proposed for query expansion to deal with ambiguity of terms. Lioma and Ounis [11] attempted two approaches for query expansion technique that is based firstly, a purely shallow syntactic-based query expansion (SSQE) technique and second, a combination of the SSQE method and the probabilistic pseudo relevance feedback approach. The SSQE method reduced the query size by retaining only sentential fragments that correspond to the part-of-speech on the basis of their high frequency. However their assumption was not accurate as frequently occurring part-of-speech blocks are merely a result of sentence construction in natural language documents. Also their approach was computationally intensive as it required parsing of documents. Another way of expanding the query which was based on co-occurrence of terms has a drawback that two terms which co-occur in the same sentence seem more correlated than two terms which occur distantly within a document but the simple co-occurrence does not necessarily mean that the terms are correlated. Moreover this approach gave more importance to rare terms than to common terms. Cao et al. [12] and Collins-Thompson and Callan [13] captured both direct and indirect term relationships for query expansion through external knowledge sources such as ontologies and statistical processing of the document corpus respectively as independent usage of the sources showed minimal improvement in retrieval performance. But there is minimal improvement as several important factors have not been examined and utilized extensively; e.g. the query structure, length, and linguistic characteristics. One of the noticeable limitations of using the WordNet Ontology given by Mihalcea [14] for query expansion is the limited coverage of concepts and phrases within the ontology.

Through this paper, a technique is derived to improve the performance of query expansion and to overcome the limitations of other approaches. The proposed approach is focused towards expanding the query by identifying relevant terms based upon weightage assigned to the similarities between the terms thus determining how similar the neighborhood terms are to the ambiguous terms, rather than considering just the co-occurrence of terms in a particular sentence. The utmost consideration is given to nouns as they contain most of the information. The given approach is appropriate for long queries also. The technique is regardless of the sentential fragments on the basis of high frequency unlike previous techniques. In this method, aggregated similarity between the every sense of the ambiguous term and the other terms in the query is found out by using the similarity measures [7] [15]. Then fuzzy weights are assigned to each similarity measure on the basis of their distance from the ambiguous term. Ordered weighted averaging (OWA) operators [16] are used to find the aggregated weights of reusable components represented by a similarity score. The sense having the maximum similarity score is considered to be the most appropriate sense for the ambiguous term. Lastly, the query is expanded by addition of those terms related to the identified sense of the ambiguous term along with other terms in the query, hence giving more relevant documents.

In section 2, the related work is mentioned, in next section, we will discuss about the basic concepts used, in section 4, the proposed method for sense resolving and query expansion is explained. Going further, in section 5, we will explain the proposed method with the help of an example along-with the results, finally in last section, we conclude our research and future work.

2. RELATED WORK

Query expansion (QE) is a technique used to improve the performance of information retrieval by addition of relevant terms into the user’s query specific to the context. In literature, different QE approaches are studied in different ways. For instance, Manning et al [1] provided a classification of QE approaches into global and local methods, where global methods are query-independent since all documents are examined for all queries. Conversely, local methods modify a query relative to the documents initially returned by the query. In Klink’s study [17], he proposed an automatic reformulation method for improving the original query wherein, the query is improved by augmenting words having similar meaning. The further categorization of QE approaches was given by Grootjen and van der Weide [18] as extensional, intentional, or collaborative ones. The first approach materializes information need in terms of documents, for instance relevance feedback and local analysis methods. The second category i.e. intentional approach is primarily thesauri / ontology-based, which take advantage of the semantics of keywords. Collaborative approaches are focused towards exploiting users’ behavior, e.g., mining query logs, as a complement to previous approaches. Bhogal et al. [19] provided a comprehensive review of ontology based query expansion, which presents several query expansion approaches, focusing on examples using corpus dependent or independent ontologies. Sanasam et al. [20] proposed a method for query expansion based on real time implicit feedback from user. Certain ambiguous terms present in the user’ query sometimes lead to the addition of irrelevant terms to the query which can inversely affect precision and retrieval. So to deal with ambiguity, huge amount of work has been done in the field of disambiguation of the words based on WordNet definitions. In Voorhees [21], author used WordNet for query expansion by adding synonyms to the query. The semantic network WordNet [22] which is a collection of words or collocations that all mean the same thing is structured as a multiple hierarchy, and its basic unit of knowledge is the synset. It can be used to determine the similarity of the words by dictionary definitions or the gloss overlaps.

Leacock [23] and Lee [24] proposed a measure of the semantic similarity by calculating the length of the path between the two nodes in the hierarchy. Heiner [25] proposed a method for determining semantic similarity by taking implicit information contained in the definitions of classes and relations. Resnik [9] used semantic similarity between two words and disambiguated noun senses. Another method is proposed by Agirre [26] which was based on the conceptual distance among the concepts in the hierarchy and compared the above systems in a real-world spelling correction system. PageRank-based WSD algorithm was also introduced by Agirre and Soroa [3], which rely not only in the semantic representation of text, but also on the PageRank formula. PageRank execution formula given by Agirre and Soroa was to concentrate the initial probability mass uniformly over the word nodes so as to constitute the context of the word to be disambiguated.

Another method which was based on the semantic distance between topics by the use of disambiguation procedure was proposed by Sussna [27]. His method used WordNet graph as a basis and examined all nouns in a window of context and each noun was assigned a sense in a way that minimizes a semantic distance function among all selected senses. Fragos [28] used WordNet glosses. Two types of bags were used to disambiguate a word: A bag of words related to every sense of the word and a bag of words related to the context. The Lesk’s approach is based on the count of the common words between the bags related to each sense and the bag related to the context. Lesk method was modified to allow for the use of any measure of semantic relatedness by Patwardhan [29]. Mihalcea [30] constructed semantic networks from WordNet and in order to address the all words task for the English language, he ran an adaptation of the PageRank algorithm on them. Methods proposed by Jiang and Conrath [31] were also based on similarity ideas. In [32], Navigli and Lapata explore several measures for analyzing the connectivity of semantic graph structures in local or global level. Banerjee and Pedersen [33] extend Lesk’s algorithm to extend the dictionary definitions with additional definitions from concepts found in the rich network of word sense relations in WordNet.

Word Sense Disambiguation (WSD) has been extremely useful in the application of natural technology using machine translation [34]. In order to addresses the problem of selecting the most appropriate sense for a word, with respect to its context was made possible by the use of word sense disambiguation as there are different senses of a word offered from a dictionary or thesaurus [3]. An unsupervised graph-based method for WSD proposed by Sinha and Mihalcea [14], was based on an algorithm that computes graph centrality of nodes in the constructed semantic graph, they made use of the in-degree, the closeness, and the betweenness of the vertices in the graph, as well as PageRank to measure the centrality of the nodes. To compute the similarity of the nodes in the semantic graph they also employ five known measures of semantic similarity or relatedness, based on an idea initially presented by Patwardhan et al. [29]. Wojtinnek et.al [35] measured the semantic relatedness of two concepts by measuring the similarity of the surroundings of their corresponding nodes in the network. Navigli and Lapata in their work [36] suggested that in order to determine the particular sense of an ambiguous word, the sense having the most number of incoming connections will be the most appropriate one. These connections can be weighted according to semantic type (e.g., synonymy relations are more important than hyponymy). They analyzed the impact of connectivity metrics for unsupervised WSD by devising a graph based algorithm. Gasperin et al [37] in their work presented syntactic information could be used to extract semantic regularities of word sequences.

3. BASIC CONCEPTS

3.1 COMPUTATIONAL LEXICONS/WORDNET

A WordNet is a word sense network. A word sense node in this network is a synset which is regarded as a basic object in the WordNet. Each synset in the WordNet is linked with other synsets through the well-known lexical and semantic relations of hypernymy, hyponymy, meronymy, troponymy, antonymy, entailment etc. Semantic relations are between synsets and lexical relations are between words. These relations serve to organize the lexical knowledge base [22].

Hyponymy and Hypernymy (is a kind of): Hypernymy is a semantic relation between two synsets to capture super-set hood. Similarly, hyponymy is a semantic relation between two synsets to capture sub-set hood. Meronymy and Holonymy (Part-whole relation): It is a semantic relation between two synsets. If the concepts *A* and *B* are related in such a manner that *A* is one of the constituent of *B*, then *A* is the meronym of *B* and *B* is the holonym of *A*.

- Entailment: Entailment refers to a relationship between two verbs. Any verb *A* entails *B*, if the truth of *B* follows logically from the truth of *A*.
- Troponymy: Troponym denotes a specific manner elaboration of another verb. It shows manner of an action, i.e., *X* is a troponym of *Y* if to *X* is to *Y* in some manner.
- Antonymy: Antonymy is a relation that holds between two words that (in a given context) express opposite meanings.
- Synonymy: Synonymy is a relation that holds between two words that express same meaning.

3.2 SIMILARITY MEASURES

The similarity between two concepts *A* and *B* is represented as *Sim* (*A*, *B*) and is related to their commonality. The more commonality they share, the more similar they are. The maximum similarity between A and B is reached when they are identical, no matter how much commonality they share. There are a number of measures that were developed to quantify the degree to which two words are semantically related using information drawn from semantic networks [14]. The various approaches that will be used in this paper are discussed below:-

3.2.1 The Leacock & Chodorow [23] Similarity Is Determined As:

$$Sim_{lch} = -\log\left(\frac{length}{2 * D}\right) \quad (1)$$

where, length is the length of the shortest path between two concepts using node-counting and D is the maximum depth of the taxonomy.

3.2.2 Another Similarity Metric Which Measures The Depth Of Two Given Concepts In The Wordnet Taxonomy Is Given By Wu And Palmer [10], Produced A Similarity Score:

$$Sim_{wup} = \frac{2 * depth(LCS)}{depth(concept\,1) + depth(concept\,2)} \quad (2)$$

3.2.3 The Similarity Between Two Concepts Can Also Be Calculated Using Shortest Distance Between The Concepts Given By [15]:

$$Sim_{s,t} = \frac{1}{dist(s,t)} \quad (3)$$

where, *s* and *t* are two concepts and distance is the shortest distance between two concepts using node counting method.

3.3 OWA OPERATORS

OWA operators [16] have been used to formalize the linguistic quantifiers. In [16], to deal with multi-criteria decision-making problems, Yager proposes a family of mean-like operators whose arguments are weighted according to their order made by sorting the arguments and then averaged according to their weights, so these operators are named ordered weighted averaging (OWA) operators. OWA operator is used to obtain the aggregated weights of the reusable components with respect to the features needed by the user by giving different weighting vectors, the OWA operators lie between the choices of the minimum and the maximum of the arguments. An OWA operator that takes n input arguments is a mapping:

$$F : R^n \rightarrow R, \quad (4)$$

which has a weighting vector W of dimension n associated with it. The weighting vector *W* has the following properties:-

$$W_j \ \varepsilon\, [0, 1] \text{ and } \sum_{i=1}^{n} w_j = 1$$

4. PROPOSED METHOD

To improve the retrieval of documents, the user query is expanded to include more relevant terms through query expansion method. As the query also contains certain ambiguous terms, it can lead to low precision results. So, in order to resolve the ambiguity of the terms present in the user's query, an approach is proposed for expanding the query by finding the appropriate sense of the ambiguous terms and then adding the related terms to the original query for efficient retrieval of documents. Through this method, a query is considered having only one ambiguous word which can have more than one sense. The other words present on both sides of an ambiguous word can be nouns, verbs, adjectives etc. but this approach is focusing only on noun form (as most of the information is represented by nouns) present on right and left side of ambiguous word. To proceed with, all the stop words and stemming are removed from the sentence, verbs, adjectives, adverbs are neglected so that only those words which are nouns including the ambiguous word are left. Then the aggregated similarity (σ) of each sense of the ambiguous word is calculated with every word present on its left and right side by averaging various similarity measures as discussed in section 3.2. The aggregated similarity is calculated as,

$$\sigma = \frac{Sim_{lch} + Sim_{wup} + Sim_{s,t}}{3} \quad (5)$$

A weighted value is then assigned to similarity measure of each term based on their distance from the ambiguous term (Fig.1) i.e. the similarity measure of the terms will be assigned higher weight if it is more close to the ambiguous term and similarity measure of the terms will be assigned lower weight if away from the ambiguous term, in such a way that the sum of the assigned weights on both side of the ambiguous term is equal. For better explanation, let us assume that the query consist of *n* terms starting from W_1, W_2, …, W_k ,…,W_n. where, W_k is the ambiguous term and $1 \leq k \leq n$.

The diagrammatic representation of weights assigned to similarity measures is given in Fig.1.

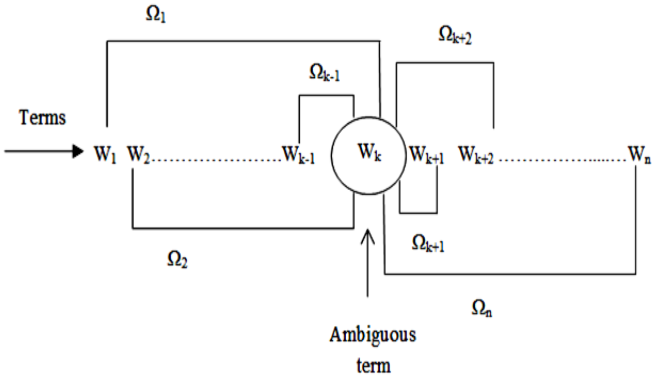
Ω1
Ωk+2
Ωk-1
Terms
W1 W2.....................Wk-1
Wk
Wk+1 Wk+2 ....Wn
Ωk+1
Ω2
Ωn
Ambiguous term

Fig.1. Diagrammatic representation of weights assigned to terms

OWA operator is then applied to obtain the similarity score of the components with respect to the features needed by the user by giving different weighting vectors, the OWA operator lies between the choices of the minimum and the maximum of the arguments (section 3.3). The sense having the highest similarity score will be selected as the most appropriate sense of the ambiguous word with respect to the context of the query and finally, query is expanded using real time implicit feedback from user which is provided at the time of search and relevant query expansion terms are determined based on the context identified [20].

4.1 THE DESCRIPTION OF THE PROPOSED APPROACH IS AS FOLLOWS

Consider a query, having verbs, nouns, adverbs, adjectives where our focus will be on nouns only. The below terms will be used further in the paper:-

(W_k) *i* be the i^{th} sense of the ambiguous word, W_k

W_k be ambiguous term in the query where $1 \leq k \leq n$

W_j be any term on LHS of W_k where $1 \leq j \leq (k\text{-}1)$

W_l be any term on RHS of W_k where $(k+1) \leq l \leq n$

σ be the aggregated similarity between the terms

Ω be the weight assigned to similarity of W_n and W_k

In this approach, a query is considered having one ambiguous term for which there are multiple senses.

1) Identifying the number of senses of the ambiguous word W_k, let's say there are 'x' senses for $(W_k)i$ where $(W_k)i$ is the i^{th} sense of W_k and $1 \leq i \leq x$.
2) The aggregated similarity 'σ' of the terms on left and right side of W_k with W_k is calculated for each of the identified sense. For each sense repeat,

   a) For LHS

   (i) Find the similarity of each term with W_k

   (ii) We derive a formula Eq.(6) to calculate the weight assigned to each similarity measure found in step (i) and it is represented by Ω_j. In order to divide the total weight equally on both the sides, we calculate a threshold value for both sides denoted by *P*, depending upon the number of terms on both sides.

   b) For RHS

   Step (i) and (ii) will be repeated to find out similarity and weighted value.

   The above algorithm can be shown with the help of a flowchart given in Fig.2.
3) With the help of OWA operator, the similarity score for current sense $(W_k)i$ is calculated by summation of product of similarity and weighted value for both LHS and RHS (calculated in step 2(a), (b)) divided by the total weight of similarity values for both sides such that the total weight is always equal to 1.
4) The sense with highest value of similarity score will be chosen as most appropriate sense of W_k.
5) The query is then expanded by addition of terms related to the identified sense of the ambiguous term. The algorithm of the proposed method is given below:

Algorithm

Step 1: Query is entered by user having an ambiguous term let's say, W_k and other non-ambiguous terms

Step 2: Get the number of senses for the ambiguous term W_k from wordNet. Let us say that there are '*X*' senses of the ambiguous term. Where, each sense can be represented as:

$(W^k)i$ represents i^{th} sense of the ambiguous term W_k where $1 < I < x$

Step 3: In order to calculate the similarity and weightage, perform the following steps:

For $i = 1$ to x

Repeat

For LHS

i) Get the similarity of all the terms on LHS of W_k with the current sense i.e $(W_k)i$

For $j = 1$ to $(k = 1)$

Get the aggregated similarity of ambiguous term with each term represented by σ_{kj}.

ii) Find the weight assigned to each similarity measure on LHS (Eq.(6)) for the current sense.

$$\Omega_j = \frac{P_{LHS}}{100} * \Omega_{k-j+1} + \frac{\sigma_{kj}}{\sum \sigma_k}\left(0.5 - \frac{P_{LHS}}{100} * \Omega_{k-j+1}\right) \qquad (6)$$

where, P_{LHS} is the threshold value and calculated as:

$$P_{LHS} = \frac{No.of\ \ tems\ on\ RHS\, of\ W_k}{total\, no.\ of\ tems\ (excluding\ ambiguous\ term)} * 100$$

And Ω_{k-j+1} is weight of the immediate right term of W_k while moving towards left.

When $j = 1$, then $\Omega_{k-j+1} = \Omega = 0.5$ (Since weight is assumed to be equally divided on both the sides of the ambiguous word W_k).

For RHS

i) Get the similarity of all the terms on RHS of W_k with the current sense i.e. $(W_k)i$

For $1 = (k+1)$ to n

Get the aggregated similarity of ambiguous term with each term represented by σ_{kj} .

ii) Find the weight assigned to each similarity measure on RHS (Eq.(7)) for the current sense.

$$\Omega_i = \frac{P_{RHS}}{100} \Omega_{k+i-1} + \frac{\sigma_{kj}}{\sum \sigma_k}\left(0.5 - \frac{P_{RHS}}{100} * \Omega_{k+i-1}\right) \qquad (7)$$

where, P_{RHS} is calculated as:

$$P_{RMS} = \frac{No.of\ \ tems\ on\ LHS\, of\ W_k}{total\, no.\ of\ tems\ (excluding\ ambiguous\ term)} * 100$$

And Ω_{k+i-1} is weight of the immediate left term of W_k while moving towards right

When $l = 1$, then $\Omega_{k+i-1} = \Omega_k = 0.5$ (since weight is assumed to be equally divided on both the sides of the ambiguous word W_k)

Step 4: Apply OWA operator to find the similarity score of current sense $(W_k)I$, using Eq.(8)

$$Similarity\ score = \left(\sum_{j=1}^{k=1} \sigma_{kj} * \Omega_j\right) + \left(\sum_{i=k+1}^{n} \sigma_{ki} * \Omega_i\right) \Bigg/ \left(\sum_{ik+1}^{n} \Omega_i + \left(\sum_{j=1}^{k-1} \Omega_j\right)\right) \qquad (8)$$

where, $\sum_{i=k+1}^{n} \Omega_i + \sum_{j=1}^{k-1} \Omega_j = 1$

Step 5: After calculating similarity score for all the senses, the most appropriate sense will be the one with highest similarity score.

Step 6: The query is expanded by taking real time implicit feedback from user [20] and adding relevant terms related to the identified sense of W_k.

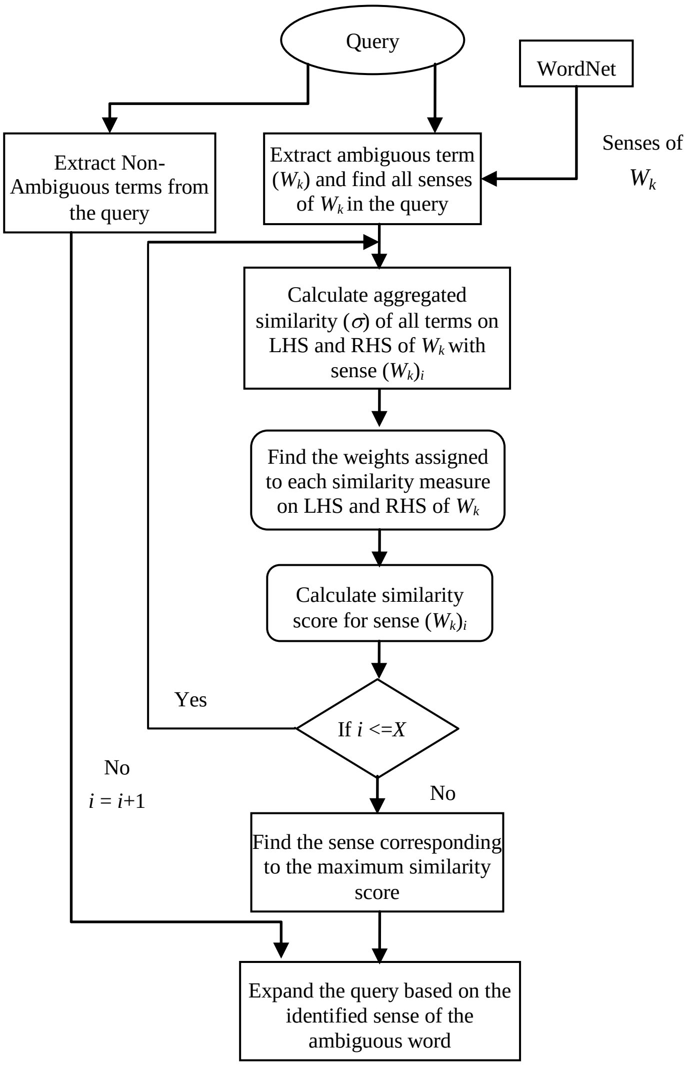
Query
WordNet
Senses of W_k
Extract Non-Ambiguous terms from the query
Extract ambiguous term (W_k) and find all senses of W_k in the query
Calculate aggregated similarity (σ) of all terms on LHS and RHS of W_k with sense $(W_k)_i$
Find the weights assigned to each similarity measure on LHS and RHS of W_k
Calculate similarity score for sense $(W_k)_i$
Yes
If i <=X
No
$i = i+1$
No
Find the sense corresponding to the maximum similarity score
Expand the query based on the identified sense of the ambiguous word

Fig.2. Flowchart

5. ILLUSTRATION THROUGH EXAMPLE AND RESULTS

The above approach can be explained through an example where a query is taken.

The sample query taken is: “He has been through hard times and without any thought of tomorrow, he lives in present and enjoys every moment of the day”.

Here except nouns which are there in the query like hard times, tomorrow, moment, day, present where “PRESENT” is the ambiguous word, all the words are neglected.

A graph is constructed around all the possible senses of the ambiguous word from WordNet and used for finding the shortest distance between a pair of concepts by node counting method. The graph is given in Fig.3

Using the graph given in Fig.3, shortest distance is found out between the terms in the query and the ambiguous term for the purpose of calculating the similarity (section 3.2). Also the WordNet noun taxonomy is used for finding out the depth of the concepts which is further used determining similarity measures.

The WordNet Noun Taxonomy for “present” is given in Fig.4:

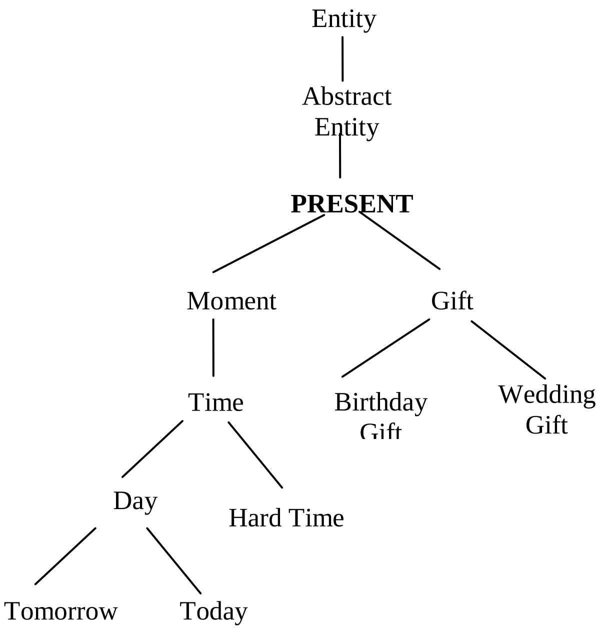

Fig.4. WordNet Noun Taxonomy

The following results are obtained (Table.1) after applying the algorithm (section 4.1). For each sense of the ambiguous term, certain parameters are evaluated namely aggregated similarity, weightage assigned to similarity values, similarity score using OWA operator and a conclusion is derived that the first noun sense of the ambiguous term present is the most appropriate sense for this ambiguous term in the query as this sense has the maximum similarity score. Thus, using this approach, the efficiency of determining the appropriate sense of an ambiguous term in the query is improved which can be further used to resolve the ambiguity in query and expand the query with appropriate and related terms only. It will result in retrieval of more relevant documents with respect to user’s query, reducing the count of irrelevant documents.

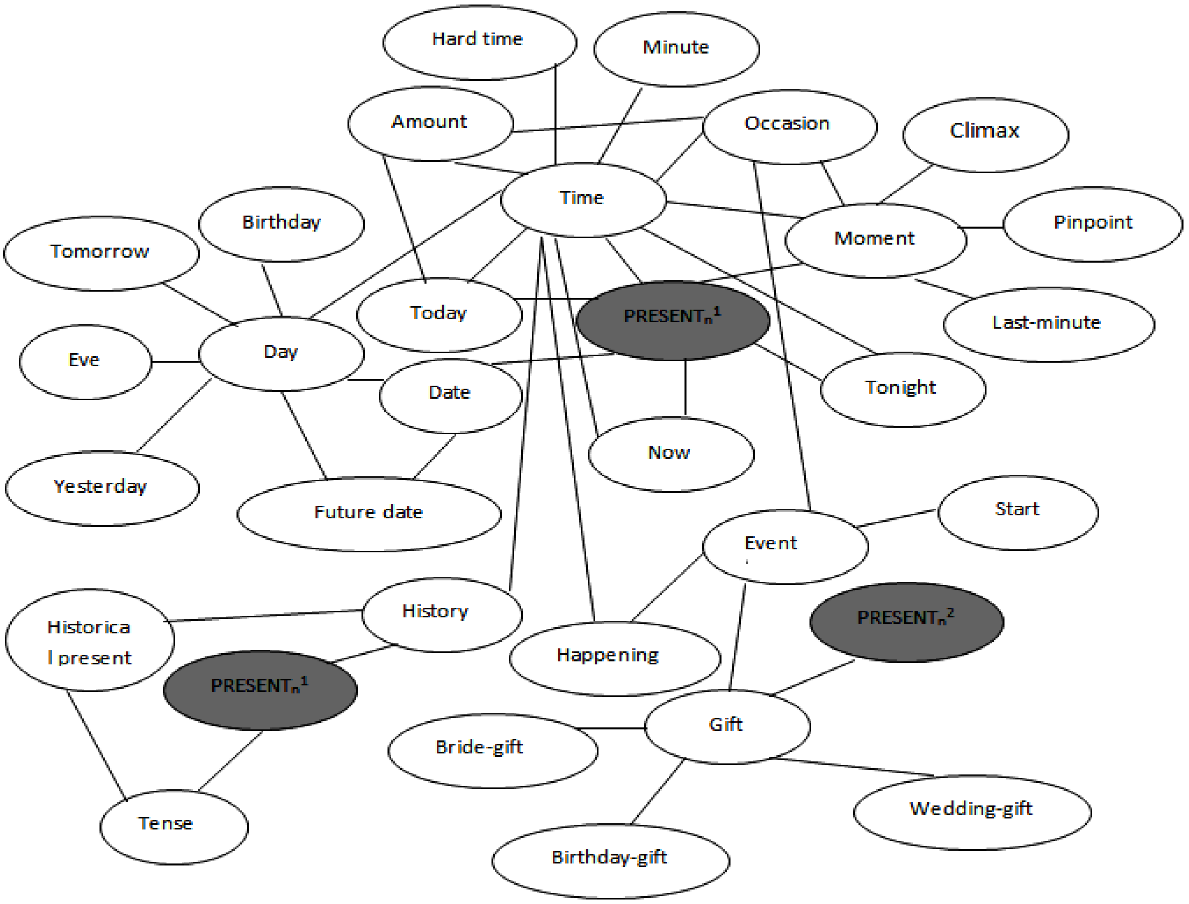

Fig3. An excerpt of WordNet around the word ’Present’

The table containing various evaluation parameters for each sense of the ambiguous word and the corresponding similarity score is given below:

Table.1. Evaluation and Results

P_{LHS} = 2/4 *100= 50% P_{RHS} = 2/4*100= 50%								
Sense	Parameters evaluated	W_1 (Hard time)	W_2 (Tomorrow)	W_k (Ambiguous word-Present)	W_3 (Moment)	W_4 (Day)	$\sum\sigma$	Similarity Score $\frac{\sum_1^n(\sigma\Omega)}{\sum_1^n(\Omega)}$
Sense 1	σ	0.52	0.39		0.83	0.52	2.26	0.57
$Present_n^1$	Ω	0.21	0.29		0.34	0.17	-	
	σ Ω	0.10	0.11		0.28	0.08	-	
Senses 2	σ	0.28	0.23		0.37	0.28	1.16	0.27
$Present_n^2$	Ω	0.21	0.29		0.32	0.18	-	
	σ Ω	0.05	0.06		0.11	0.05	-	
Sense 3	σ	0.40	0.32		0.44	0.40	1.54	0.38
$Present_n^3$	Ω	0.20	0.30		0.32	0.18	-	
	σ Ω	0.08	0.09		0.14	0.07	-	

6. CONCLUSION

This paper presented a method of improving query expansion by resolving the sense of ambiguous terms present in the query. The semantic similarity or relatedness between the ambiguous term and other terms is found out using one or more similarity measures. A weighted value is then is assigned to similarity measure which is calculated for every term present on both the sides of ambiguous word, based on the distance from the ambiguous term. OWA operator is used to find out aggregated similarity score corresponding to each sense and the sense having highest similarity score is considered to be the most appropriate sense for the particular term. After resolving the sense of the ambiguous word, the query is expanded using real time implicit feedback from user so as to retrieve more relevant documents with respect to the context of the keywords in the query and improve the retrieval efficiency.

In future, the work can be expanded for all types of words like verbs, adverbs, adjectives etc. along with nouns and try to resolve the sense of the ambiguous word present in the query and further expand the query

REFERENCES

[1] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze, "*An Introduction to Information Retrieval*", Cambridge University Press, 2009.

[2] Heiner Stuckenschmidt, "Data Semantics on the Web", *Journal on Data Semantics*, pp. 1-9, 2012.

[3] George Tsatsaronis, Iraklis Varlamis and Kjetil Nørvag, "An Experimental Study on Unsupervised graph-based Word Sense Disambiguation", *Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science*, Vol. 6008, pp. 184-198, 2010.

[4] Roy Rada, Hafedh Mili, Ellen Bicknell and Maria Blettner, "Development and Application of a Metric on Semantic Nets", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 19, No. 1, pp. 17-30, 1989.

[5] Michael Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone", *Proceedings of 5th Annual International Conference on Systems Documentation-SIGDOC*, pp. 24-26, 1986.

[6] Christiane Fellbaum, "*WordNet: An Electronic Lexical Database*", Bradford books, 1998.

[7] Sukanya Manna and B. Sumudu U. Mendis, "Fuzzy Word Similarity: A Semantic Approach Using WordNet", *IEEE International Conference on Fuzzy Systems*, pp. 1-8, 2010.

[8] M. Ross Quillian, "Word concepts: A theory and simulation of some basic semantic capabilities", *Behavioural Science*, Vol. 12, No. 5, pp. 410-430, 1967.

[9] Philip Resnik, "Using information content to evaluate semantic similarity in a taxonomy", *Proceedings of the 14th International Joint Conference on Artificial intelligence*, Vol. 1, pp. 448-453, 1995.

[10] Z. Wu and M. Palmer, "Verb semantics and lexical selection", *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, pp. 133-138, 1994.

[11] C. Lioma, and I. Ounis, "A syntactically-based query reformulation technique for information retrieval", *Information Processing and Management*, Vol. 44, No. 1, pp. 143-162, 2008.

[12] Guihong Cao, Jian-Yun Nie and Jing Bai, "Integrating word relationships into language models", *Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 298-305, 2005.

[13] Kevyn Collins-Thompson and Jamie Callan, "Query expansion using random walk models", *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pp. 704-711, 2005

[14] Ravi Sinha and Rada Mihalcea, "Unsupervised graph-based word sense disambiguation using measures of word semantic similarity", *Proceedings of the International Conference on Semantic Computing*, pp. 363-369, 2007.

[15] Troy Simpson and Thanh Dao, "WordNet based Semantic Similarity Measurement", http://www.codeproject.com/Articles/11835/WordNet-based-semantic-similarity-measurement, 2010.

[16] R. R. Yager, "On ordered weighted averaging aggregation operators in multi-criteria decision making", *IEEE Transactions on Systems, Man and Cybernetics*, pp. 183-190, 1988.

[17] Stefan Klink, "Query reformulation with collaborative concept-based expansion", *Proceedings of the First International Workshop on Web Document Analysis*, 2001.

[18] F. A. Grootjen and Th. P. Van Der Weide, "Conceptual query expansion. Data Knowledge", *Data and Knowledge Engineering*, Vol. 56, pp. 174-193, 2006.

[19] J. Bhogal, A. Macfarlane, and P. Smith, "A review of ontology based query expansion", *Information Processing and Management*, Vol. 43, No. 4, pp. 866-886, 2007.

[20] Sanasam R. Singh, Hema A. Murthy and Timothy A. Gonsalves, "Inference based Query Expansion Using User's Real Time Implicit Feedback", *Knowledge Engineering and Knowledge Management, Communications in Computer and Information Science*, Vol. 272, pp. 158-172, 2013.

[21] Ellen M. Voorhees, "Query expansion using lexical-semantic relations", *Proceedings of the 17th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 61-69, 1994.

[22] George A. Miller, "WordNet: A lexical database for English", *Communications of the ACM*, Vol. 38, No. 11, 1995.

[23] C. Fellbaum and G. Miller, "Combining local context and WordNet similarity for word sense identification", *WordNet: An Electronic Lexical Database*, pp. 265-283, 1998.

[24] Joon Ho Lee, Myoung Ho Kim and Yoon Joon Lee, "Information retrieval based on conceptual distance in IS-A

hierarchies”, *Journal of Documentation*, Vol. 49, No. 2, pp. 188-207, 1993.

[25] Heiner-Stuckenschmidt, “A Semantic Similarity Measure for Ontology-Based Information”, *Flexible Query Answering Systems Lecture Notes in Computer Science*, Vol. 5822, pp. 406-417, 2009.

[26] Eneko Agirre and German Rigau, “Word Sense Disambiguation Using Conceptual Density”, Proceedings *of the 16th Conference on Computational Linguistics*, Vol. 1, 1996.

[27] Michael Sussna, “Word sense disambiguation for free-text indexing using a massive semantic network”, Proceedings *of the Second International Conference on Information and Knowledge Management*, pp. 67-74, 1993.

[28] Kostas Fragos, Yannis Maistros and Christos Skourlas, “Word-sense disambiguation using WordNet relations”, *Proceedings of the 1st Balkan Conference in Informatics*, 2003.

[29] Siddharth Patwardhan, Satanjeev Banerjee and Ted Pedersen, “Using measures of semantic relatedness for word sense disambiguation”, *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 241-257, 2003.

[30] Rada Mihalcea, Paul Tarau and Elizabeth Figa, “PageRank on semantic networks, with application to word sense disambiguation”, *Proceedings of the 20th International Conference on Computational Linguistics*, 2004.

[31] Jay J. Jiang and David W. Conrath, “Semantic similarity based on corpus statistics and lexical taxonomy”, *Proceedings of International Conference Research on Computational Linguistics*, pp. 19-33, 1997.

[32] Roberto Navigli and Mirella Lapata, “Graph connectivity measures for unsupervised word sense disambiguation”, *Proceedings of the 20th International Joint Conference on Artificial intelligence*, pp. 1683-1688, 2007.

[33] Satanjeev Banerjee and Ted Pedersen, “An adapted Lesk algorithm for word sense disambiguation using WordNet”, *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 136-145, 2002.

[34] Marine Carpuat and Dekai Wu, “Improving statistical machine translation using word sense disambiguation”, *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 61-72, 2007.

[35] Pia-Ramona Wojtinnek, Stephen Pulman and Johanna Völker, “Building Semantic Networks from Plain Text and Wikipedia with Application to Semantic Relatedness and Noun Compound Paraphrasing”, *International Journal of Semantic Computing, World Scientific*, Vol. 6, No. 1, pp. 1-25, 2012

[36] R. Navigli and M. Lapata, “An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation”, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 4, pp. 678-692, 2010.

[37] Caroline Gasperin, Pablo Gamallo, Alexandre Agustini, Gabriel Lopes, Vera De Lima and Faculdade De Informtica, “Using syntactic contexts for measuring word similarity”, *Proceedings of the Workshop on Semantic Knowledge Acquisition and Categorisation*, 2001.