A New Semantic Similarity Metric for Solving Sparse Data Problem in Ontology based Information Retrieval System

K.Saruladha[1], Dr.G.Aghila[2], Sajina Raj[3]

[1]Research Scholar,
Department of Computer Science & Engg.
Pondicherry Engineering College,
Pondicherry 605 014, India.

[2]Department of Computer Science,
Pondicherry University,
Pondicherry 605 014, India.

[3]Post Graduate Student,
Pondicherry Engineering College,
Pondicherry 605 014, India.

Abstract

Semantic similarity assessing methods play a central role in many research areas such as Psychology, cognitive science, information retrieval biomedicine and Artificial intelligence. This paper discuss the existing semantic similarity assessing methods and identify how these could be exploited to calculate accurately the semantic similarity of WordNet concepts. The semantic similarity approaches could broadly be classified into three different categories: Ontology based approaches (structural approach), information theoretic approaches (corpus based approach) and hybrid approaches. All of these similarity measures are expected to preferably adhere to certain basic properties of information. The survey revealed the following drawbacks The information theoretic measures are dependent on the corpus and the presence or absence of a concept in the corpus affects the information content metric. For the concepts not present in the corpus the value of information content tends to become zero or infinity and hence the semantic similarity measure calculated based on this metric do not reflect the actual information content of the concept. Hence in this paper we propose a new information content metric which provides a solution to the sparse data problem prevalent in corpus based approaches. The proposed measure is corpus independent and takes into consideration hyponomy and meronomy relations. Empirical studies of finding similarity of R&G data set using existing Resnik, lin and J& C semantic similarity methods with the proposed information content metric is to be studied. We also propose a new semantic similarity measure based on the proposed information content metric and hypernym relations.

The correctness of the information content metric proposed is to be proved by comparing the results against the human judgments available for R& G set. Further the information content metric used earlier by Resnik, lin and Jiang and Cornath methods may produce better results with alternate corpora other than brown corpus. Hence the effect of corpus based information content metric on alternate corpora is also investigated.

Keywords-Ontology, similarity method, information retrieval, conceptual similarity, taxonomy, corpus based

1 . INTRODUCTION

The goal of Information retrieval process is to retrieve Information relevant to a given request. The aim is to retrieve all the relevant information eliminating the non-relevant information. An information retrieval system comprises of representation, semantic similarity matching function and Query. Representation comprises the abstract description of documents in the system. The semantic similarity matching function defines how to compare query requests to the stored descriptions in the representation.

The percentage of relevant information we get mainly depends on the semantic similarity matching function we used. So far, there are several semantic similarity methods used which have certain limitations despite the advantages. No one method replaces all the semantic similarity methods. When a new information retrieval system is going to be build, several questions arises related to the semantic similarity matching function to be used. In the last few decades, the number of semantic similarity methods developed is high.

This paper discusses the overall view of different similarity measuring methods used to compare and find very similar concepts of ontology. We also discuss about the pros and cons of existing similarity metrics. We have presented a new approach which is independent of the corpora for finding the semantic similarity between two concepts. Section II, a set of basic intuitive properties are defined to which the compatibility of similarity measures in information is preferable. Section III discusses various approaches used for similarity computation and the limitations of those methods. In Section IV, comparison among different semantic similarity measures are discussed. In Section V we introduce a new semantic similarity approach and an algorithm for finding the similarity between all the relations in the WordNet taxonomy. The results based on new similarity measure is promising.

2. ONTOLOGY SIMILARITY

In this section, a set of intuitive and qualitative properties that a similarity method should adhere to is discussed.[20]

- *Basic Properties*

  Any similarity measure must be compatible with the basic properties as they express the exact notion of property.

  - *CommonalityProperty*
  - *Difference Property*
  - *Identity Property*

- *Retrieval Specific Properties*

  The similarity measure cannot be symmetrical in case of ontology based information retrieval context. The similarity is directly proportional to specialization and inversely proportional to generalization.

  - Generalization Property

- *Structure Specific Properties*

  The distance represented by an edge should be reduced with an increasing depth.

  - *Depth Property*
  - *Multiple Paths Property*

3. APPROACHES USED FOR SIMILARITY COMPUTATION

In this section, we discuss about various similarity methods[20]. The similarity methods are

- *Path Length Approaches*
- *Depth-Relative Approaches*
- *Corpus-based Approaches*
- *Multiple-Paths Approaches*

3.1 Path Length Approach

The shortest path length and the weighted shortest path are the two taxonomy based approaches for measuring similarity through inclusion relation.

Shortest Path Length

A simple way to measure semantic similarity in a taxonomy is to evaluate the distance between the nodes corresponding to the items being compared. The shorter distance results in high similarity.

In Rada et al. [1989][1][14], shortest path length approach is followed assuming that the number of edges between terms in a taxonomy is a measure of conceptual distance between concepts.

distRada(*ci; cj*) = minimal number of edges in a path from *ci* to *cj*

This method yields good results. since the paths are restricted to ISA relation, the path lengths corresponds to conceptual distance. Moreover, the experiment has been conducted for specific domain ensuring the hierarchical homogeneity.

The drawback with this approach is that, it is compatible only with commonality and difference properties and not with identity property.

3.1.2. Weighted Shortest Path Length

This is another simple edge-counting approach. In this method, weights are assigned to edges. In brief, weighted shortest path measure is a generalization of the shortest path length. Obviously it supports commonality and difference properties.

- Similarity of immediate specialisation
-Similarity of immediate generalisation

$P=(p_1,\ldots..,p_n)$

where,

P_i ISA p_{i+1} or P_{i+1} ISA p_i
For each I with $x=p_1$ and $y=p_n$

Given a path $P=(p_1,\ldots..p_n)$, set s(P) to the number of specializations and g(P) to the number of generalizations along the path P as follows:

$$s(P)= |\{i\backslash p_i \text{ ISA } p_{i+1}\}| \quad (1)$$

$$g(P)=|\{i|P_{i+1} \text{ ISA } p_i\}| \quad (2)$$

If $p_1,\ldots\ldots p_m$ are all paths connecting x and y, then the degree to which y is similar to x can be defined as follows:

$$\text{sim}_{WSP}(x,y)=\max_{j=1,\ldots.m}\{ {}^{s(pj)} \quad {}^{s(pj)}\}(3)$$

The similarity between two concepts x and y, sim(x,y) WSP(weighted Shortest Path) is calculated as the maximal product of weights along the paths between *x* and *y*. Similarity can be derived as the products of weights on the paths.

$$^{g(Pj} \quad \sigma \in [0,1] \text{ and } \gamma \in [0,1] \max{}^{s(pj)}$$

s(P^j) and g(P^j) = 0

Hence the weighted shortest path length overcomes the limitations of shortest path length wherein the measure is based on generalization property and achieves identity property.

3.2 Depth-Relative Approaches

Even though the edge counting method is simple, it limits the representation of uniform distances on the edges in the taxonomy. This approach supports structure specific property as the distance represented by an edge should be reduced with an increasing depth.

3.2.1 Depth Relative Scaling

In his depth-relative scaling approach Sussna[1993][2] defines two edges representing inverse relations for each edge in a taxonomy. The weight attached to each relation *r* is a value in the range [min*r*; max*r*]. The point in the range for a relation *r* from concept *c*1 to *c*2 depends on the number *nr* of edges of the same type, leaving *c*1, which is denoted as the type specifc fanout factor:

$$W(c_{1\rightarrow r} c_2)=max_r-\{max_{r-}min_r/n_r(c_1)\}$$

The two inverse weights are averaged and scaled by depth *d* of the edge in the overall taxonomy. The distance between adjacent nodes *c*1 and *c*2 are computed as:

$$dist_{sussna}(c_1,c_2)=w(c_{1\rightarrow r} c_2)+ (c_{1\rightarrow r'} c_2)/2d \qquad (4)$$

where *r* is the relation that holds between *c*1 and *c*2, and *r'* is its inverse. The semantic distance between two arbitrary concepts *c*1 and *c*2 is computed as the sum of distances between the pairs of adjacent concepts along the shortest path connecting *c*1 and *c*2.

3.2.2 Conceptual Similarity

Wu and Palmer [1994][3], propose a measure of semantic similarity on the semantic representation of verbs in computer systems and its impact on lexical selection problems in machine translation. Wu and Palmer define *conceptual similarity* between a pair of concepts *c*1 and *c*2 as:

$$Sim_{wu\&palmer}(c_1,c_2)= \quad 2*N3/(\qquad (5)$$

Where N1 is the number of nodes on the path from c1 to a concept c3. , denoting the least upper bound of both *c*1 and *c*2. *N*2 is the number of nodes on a path from *c*2 to *c*3. *N*3 is the number of nodes from *c*3 to the most general concept.

3.2.3 Normalised Path Length

Leacock and Chodorow [1998][4], proposed an approach for measuring semantic similarity as the shortest path using is a hierarchies for nouns in WordNet. This measure determines the semantic similarity between two synsets (concepts) by finding the shortest path and by scaling using the depth of the taxonomy:

$$Sim_{Leacock\&Chaodorow}(c_1,c_2)= -log(N_p(c_1,c_2)/2D) \qquad (6)$$

Np (*c*1,*c*2) denotes the shortest path between the synsets (measured in nodes), and *D* is the maximum depth of the taxonomy.

3.3 Corpus-based Approach

The knowledge disclosed by the corpus analysis is used to intensify the information already present in the ontologies or taxonomies. In this method, presents three approaches that incorporate corpus analysis as an additional, and qualitatively different knowledge source.

3.3.1 Information Content

In this method rather than counting edges in the shortest path, they select the maximum information content of the least upper bound between two concepts. Resnik [1999] [5], argued that a widely acknowledged problem with edge-counting approaches was that they typically rely on the notion that edges represent uniform distances. According to Resnik's measure, information content, uses knowledge from a corpus about the use of senses to express non-uniform distances.

Let *C* denote the set of concepts in a taxonomy that permits multiple inheritance and associates with each concept *c* 2 *C*, the probability *p*(*c*) of encountering an instance of concept *c*. For a pair of concepts *c*1 and *c*2, their similarity can be defined as:

$$Sim_{Resnik} \quad = \max_{[illegible]} \qquad (7)$$

where,

S(*c*1,*c*2): Set of least upper bounds in the taxonomy of *c*1 and c2

p(*c*) :Monotonically non-decreasing as one moves up in the taxonomy,

$$p(c1) \leq p(c2), \text{ if c1 is a c2.}$$

The similarity between the two words w1 and w2 can be computed as:

$$wsim_{Resnik} \quad (w1, w2) = \max_{c1 \in s(w1), c2 \in s(w2)} [sim \qquad (8)$$

Where,

s(*wi*): Set of possible senses for the word *wi*.

Resnik describes an implementation based on *information content* using WordNet's [Miller, 1990][6], taxonomy of noun concepts [1999]. The information content of each concept is calculated using noun frequencies

$$Freq(c) = \quad \Sigma_n$$

where,

words(*c*): Set of words whose senses are subsumed by concept c.

$$(c)=freq(c)/N$$

where ,

N: is the total number of nouns.

The major drawback of the information content approach is that they fail to comply with the generalization property, due to symmetry.

3.3.2 Jiang and Conrath's Approach(Hybrid Method)

Jiang and Conrath [1997][7]proposed a method to synthesize edge-counting methods and information content into a combined model by adding the information content as a corrective factor.

The edge weight between a child concept *cc* and a parent concept *cp* can be calculated by considering factors such as local density in the taxonomy, node depth, and link type as,

$Wt(c_c,c_p)=$

$$\left(\beta + (1-\beta)\frac{\bar{E}}{E(cp)}\right)\left(\frac{d(cp)+1}{d(cp)}\right)^{\alpha} LS(cc,cp)T(cc, \quad (9)$$

Where,

d(*cp*) : Depth of the concept *cp* in the taxonomy,
E(*cp*) :Number of children of *cp* (the local density)
(^1E) : Average density in the entire taxonomy,
LS(*cc, cp*) : Strength of the edge between *cc* and *cp*,
T(*cc,cp*) : Edge relation/type factor

The parameters 0 and *,0* 1 control the in°uence of concept depth and density, respectively. strength of a link *LS*(*cc; cp*) between parent and child concepts is proportional to the conditional probability *p*(*cc*|*cp*) of encountering an instance of the child concept, *cc*, given an instance of the parent concept, *cp*:

$$LS(cc, cp) = -\log p(cc|cp)$$

Resnik assigned probability to the concepts as *p*(*cc* *cp*) = *p*(*cc*), because any instance of a child concept *cc* is also an instance of the parent concept *cp*. Then:

$p(C_c|C_p) =$

$p(C_c|C_p) =$

If *IC*(*c*) denotes the information content of concept *c*. then:

$$LS(C_c,C_p) = IC(C_c)-IC(C_p)$$

Jiang and Conrath then defined the semantic distance between two nodes as the summation of edge weights along the shortest path between them [J. Jiang, 1997]:

$dist_{jiang\&conrath}(C_1,C_2)=$

$$\sum_{c\in\{path(c1,c2)-LSuper(c1,c2)\}} wt(c, pare \quad (10)$$

Where,

path(*c*1, *c*2) : the set of all nodes along the shortest path between c1 and c2
parent(*c*) : is the parent node of *c*
LSuper(*c*1, *c*2) : is the lowest superordinate (least upper bound) on the path between *c*1 *and c2..*

Jiang and Conrath's approach made *information content compatible with* the basic properties and the depth property, but not to the generalization property.

3.3.3 Lin's Universal Similarity Measure

Lin [1997; 1998][8][13] defines a measure of similarity claimed to be both universally applicable to arbitrary objects and theoretically justified. He achieved generality from a set of assumptions.

Lin's information-theoretic definition of similarity builds on three basic properties, commonality, difference and identity. In addition to these properties he assumed that the commonality between *A* and *B* is measured by the amount of information contained in the proposition that states the commonalities between them, formally:

$$I(common(A,B)) = -\log p(common(A,B)$$

where,

I(*s*): Negative logarithm of the probability of the proposition, as described by Shannon[1949].

The difference between *A* and *B* is measured by:

$$I(description(A,B)) – I(common(A,B))$$

Where,

description(*A;B*) : Proposition about what *A* and *B* are.

Lin proved that the similarity between *A* and *B* is measured by,

$$sim_{Lin}(A,B)= \quad (11)$$

The ratio between the amount of information needed to state the commonality of *A* and *B* and the information needed to describe them fully.

Lin's similarity between two concepts in a taxonomy ensures that:

$$Sim_{Lin(c1,c2)}= \quad (12)$$

where,

LUB(*c*1, *c*2): Least upper bound of *c*1 and *c*2
p(*x*) : Estimated based on statistics from a sense tagged corpus.

This approach comply with the set of basic properties and the depth property, but would fail to comply with the generalization property as it is symmetric.

3.4 Multiple-Paths Approaches

This approach solves the problem with single path approach. Single path as a measure for the similarity, fails to truly express similarity whenever the ontologies allow multiple inheritance. In multiple-path approach measurement is made by taking into account all the semantic relations in ontologies, considering more than one path between concepts. Attributes should influence the measure of similarity, thus allowing two concepts sharing the same attribute to be considered as more similar, compared to concepts not having this particular attribute.

3.4.1 Medium-Strong Relations

Hirst and St-Onge [Hirst and St-Onge, 1998; St-Onge, 1995] [9][15], distinguishes the nouns in the Wordnet as extra-strong, strong and medium-strong relations. The extra-strong relation is only between a word and its literal repetition.

3.4.2 Generalised Weighted Shortest Path

The principle of weighted path similarity can be generalized by introducing similarity factors for the semantic relations. However, there does not seem to be an obvious way to differentiate based on direction. Thus, we can generalize simply by introducing a single similarity factor and simplify to bidirectional edges. This method solves the symmetry problem by introducing weighted edges.

3.4.3 Shared Nodes

This approach overcomes the limitation of single path length approach. Multiple paths are considered for measuring the similarity.

The shared nodes approach with similarity function discussed above complies with all the defined properties.

3.4.4 Weighted Shared Nodes Similarity

It is found that when deriving similarity using the notion of shared nodes, not all nodes are equally important.

Assigning weights to edges is very important, as it generalizes the measure so that it can be make use for different domains with different semantic relations. The weighted shared nodes measure complies with all the defined properties.

4. COMPARISON OF DIFFERENT SIMILARITY MEASURES

In this section we discuss about the results of comparison of the measures to human similarity judgments.

Table 1: Replica Of The Rubinstein And Goodenough And The Miller And Charles Experiments

Word1	Word2	Replica	R&G	M&C
Car	Automobile	3.82	3.92	3.92
Gem	Jewel	3.86	3.84	3.94
Journey	Voyage	3.58	3.54	3.58
Boy	Lad	3.10	3.76	3.84
Coast	Shore	3.38	3.70	3.60
Asylum	madhouse	2.14	3.61	3.04
Magician	Wizard	3.68	3.50	3.21
Midday	Noon	3.45	3.42	3.94
Furnace	Stove	2.60	3.11	3.11
Food	Fruit	2.87	3.08	2.69
Bird	Cock	2.62	3.05	2.63
Bird	Crane	2.08	2.97	2.63
Tool	implement	1.70	2.95	3.66
Brother	Monk	2.38	2.82	2.74
Lad	Brother	1.39	1.66	2.41
Crane	Implement	1.26	1.68	2.37
Journey	Car	1.05	1.16	1.55
Monk	Oracle	0.90	1.10	0.91
Cemetery	Woodland	0.32	0.95	1.18
Food	Rooster	1.18	0.89	1.09
Coast	Hill	1.24	0.87	1.26
Forest	Graveyard	0.41	0.84	1.00
Shore	Woodland	0.81	0.63	0.90
Monk	Slave	0.36	0.55	0.57
Coast	Forest	0.70	0.42	0.85
Lad	Wizard	0.61	0.42	0.99
Chord	Smile	0.15	0.13	0.02
Glass	Magician	0.52	0.11	0.44
Rooster	Voyage	0.02	0.08	0.04
Noon	String	0.02	0.08	0.04

The first human similarity judgment was done by Rubinstein and Goodenough [1965][11], using two groups totaling 51 subjects to perform synonymy judgments on 65 pairs of nouns and this in turn been the basis of the comparison of similarity measures.

Miller and Charles [1991][12] repeated Rubinstein and Goodenough's original experiment, they used a subset of 30 noun pairs from the original list of 65 pairs, where ten pairs were from the high level of synonymy, ten from the middle level and ten from the low level.

The correlation between these two experiments is 0.97.

An experiment were performed by taking the r*eplica* of the Miller and Charles experiment that included 30 concept pairs and an additional ten new compound concepts pairs and human judgment for all the 40 pairs has been performed. The correlations for the measures done by Resnik, Jiang and Conrath; Lin, Hirst and St-Onge; and Leacock and Chodorow are shown in table given below. [Budanitsky, 2001][18][19][20]

Table 2: Correlation between Different Similarity Measures & Human Similarity Judgments from the Miller and Charles Experiment

Approach	Correlation
Resnik	0.744
Jiang and Conrath	0.850
Lin	0.829
Hirst and St-Onge	0.744
Leacock and Chodorow	0.816

The table given below shows the correlations between the replica and the two previous experiments. The correlations between the replica experiment and the previous experiments are fairly good[20].

Table 3: Correlation between the Three Human Similarity Judgment Experiments

Correlation		
Rubinstein Goodenough	Miller & Charles	0.97
Rubinstein & Goodenough	Replica	0.93
Miller & Charles	Replica	0.95

4.1 Sparse Data Problem in Corpus based Approaches

Following the standard definition from Shannon and Weaver's information theory [1949], the information content of concept c is $-\log p(c)$.

Information content in the context of WordNet is drawn from the Brown university standard corpus This corpus refers to a collection of documents of widely varying genres collected in 1961 which was updated in 1971and 1979 to reflect new literatures shown in table 4 The collection of more than 1 million words was manually tagged with about 80 parts of speech. The presently available list of WordNet concepts tagged in the brown corpus includes approximately 420000 words. There are more number of concepts in WordNet which are not tagged in Brown Corpus.

Table 4: Brown corpus categories

Press : reportage (Political Sports,Society,Spot News,Financial,Cultural
Press: Editorial (Institutional Daily, Personal, Letters to the Editor)
Press: Reviews (Theatre, Books, Music, Dance)
Religion (Books, Periodicals, Tracts)
Skill and Hobbies (Books, Periodicals)
Popular Lore (Books, Periodicals)
Belles-Lettres (Books, Periodicals)
Miscellaneous: US Government & House Organs (Government Documents, Foundation Reports, College Catalog, Industry House organ)
Learned (Natural Sciences, Medicine, Mathematics, Social and Behavioral Sciences, Political Science, Law, Education, Humanities, Technology and Engineering)

We ran a small experiment on the Brown corpus and found that the words soccer, fruitcake, world, trade centre were not not found in the corpus.

Table 5: Incompleteness of Brown Corpus

Nouns which are not present in Brown Corpus
Soccer
Fruitcake
CPU
Autograph
Serf
Slave

The IC value tends to become zero for concepts not existing in the corpora as –log(p(c)) formula is used for IC calculation..

4.2 Significanse of the work

We have proposed an algorithm based on new information content which takes into consideration the meronomy relations. The Information content metric proposed by Pierro [2009] takes into concern the holonym relations that a concept has with other concepts.. All of the concepts of the WordNet will not have all of the relation types. Some of the concepts will have holomy relation and few with meronomy relations etc. The information content metric if it ignores the other type of relations and if it produces a zero value it means that the concept has no information which is not correct. Hence we decided to consider both the meronym and holonym relations. If a concept do not have holonymy relation then the information imparted by meronomy is taken into consideration and the information content will become zero when the concept has no relations with other concepts which is a very rare case.

5. PROPOSED WORK

The semantic similarity measures are mostly based on the information content.

Most of the corpus based similarity methods like Lin[13], Jiang Cornath[2] and resnik[11] are IC based and the IC calculation is done using Brown corpus. Brown corpus suffers from the following drawbacks. The Parts of speech tags from the Brown corpus does not correspond to the hierarchical structure of concepts in WordNet. All concepts of WordNet are not present in the Brown Corpus. The noun such as autograph, serf and slave are not present in the Brown Corpus.

Similarity measures that rely on information content can produce a zero value for even the most intuitive pairs because the majority of WordNet concepts occur with a frequency of zero. This makes the Lin method and Jiang cornath method to return zero or infinity in the continuous domain and hence the similarity measure is not true or reliable.

Hence the computation of Information content should be computed in a different way so that the similarity measure becomes reliable.

5.1 Proposed Corpora Independent Similarity Measure to Solve Sparse Data Problem

The objective of this work is twofold.

1. To design a new information metric which solves sparse data problem which is independent of the corpora and which is based on semantic relations available in WordNet.

2. To investigate the existing IC metric on corpora other than brown corpus.

Table 5: Relations defined in wordnet taxonomy

Relation	Example
Meronym(Part-of)	Engine is a Meronym of Car
Holonym(Has-part)	Car is a Holonym of Engine
Meronym(Has-Member)	Team has member player
Holonym (Member-of)	Player is the member of Team

The Existing Similarity methods do not consider the Holonymy/Meronymy relationships defined in Wordnet. Hence we propose to devise a new similarity measure which considers these relationships and experimentally evaluate and compare it with the existing similarity measures using R&G data set and extended data set.

The fig.1 shows the semantic similarity system architecture.

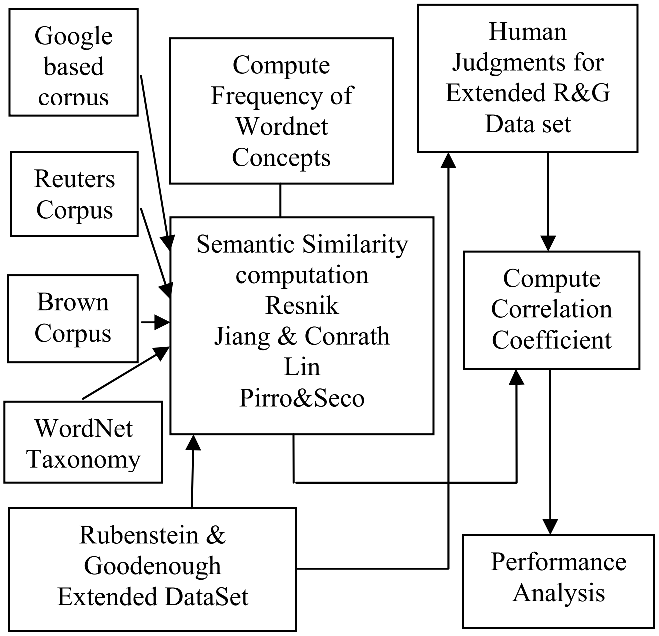
Google based corpus
Reuters Corpus
Brown Corpus
WordNet Taxonomy
Compute Frequency of Wordnet Concepts
Semantic Similarity computation Resnik Jiang & Conrath Lin Pirro&Seco
Human Judgments for Extended R&G Data set
Compute Correlation Coefficient
Rubenstein & Goodenough Extended DataSet
Performance Analysis

Fig. 1 Semantic Similarity System Architecture

5.2 Proposed IC Calculation Involving Meronym

The intrinsic IC for a concept c is defined as:

$$\mathbf{IC(c) = 1- \frac{log(hypo(c)+mero(c)+1}{Log(max_{con})}} \quad \text{---(16)}$$

where,

Hypo(c)-Function returns the number of hyponyms
Mero(c)-Function returns the number of meronyms
Max$_{con}$- is a constant that indicates the total number of concepts in the considered taxonomy
Note: Assume hypo(c), mero(c)>=0 & max$_{con}$>0

The function hypo and mero returns the number of hyponyms and meronyms of a given concept c. Note that concepts representing leaves in the taxonomy will have an IC of one, since they do not have hyponyms. The value of one states that a concept is maximally expressed and cannot be further differentiated. Moreover maxcon is a constant that indicates the total number of concepts in the considered taxonomy.

5.3 PROPOSED SIMILARITY FUNCTION BASED ON PROPOSED IC

According to our new formula, the similarity between the two concepts c1 and c2 can be defined as,

$$\mathbf{Sim_{ext}(c1,c2) =}$$

$$\mathbf{IC\ (LCS\ (c1,c2))- \sum_{hyper(c1,c2)}^{LCS(c1,c2)} hyper(c1) + hyper(c2) - hyper\ (c1 \cap c2)} \quad \text{----(17)}$$

where, the function LCS finds the lowest common subsumer of the two concepts c1 and c2 and the function hyper finds all the hypernyms of c1 and c2 upto the LCS node.

The Proposed IC formula will be used in existing semantic similarity methods such as Resnik, Lin and Jiang and cornath for computation of similarity measure suing R&G and M&C data sets. The influence of meronomy and hyponomy relations in calculation of similarity measure will be studied. The proposed similarity function will be tested with R&G and M&C data sets.

The proposed similarity function with noun pairs of R&G data set was tested for Resnik method and the correlation coefficient with human judgment were calculated. The results were promising. We have to test the proposed similarity measure against other data set.

Most of the IC based approaches are tested against the well known corpus. The concepts not present in Brown

Input (noun1, noun2), Output (similarity score)

Step 1: For each sense c1 and c2 in noun1 and noun2
If c1 and c2 are hyponyms

5.4 Proposed Algorithm

$$\mathbf{IC(c) = 1- \frac{log(hypo(c)+1)}{log(max_{con})}}$$

go to step 3

Else if c1 and c2 are meronyms

Calculate IC(c1) for meronyms i.e

$$\mathbf{IC(c) = 1- \frac{log(mero(c)+1)}{log(max_{con})}}$$

go to step 3

Step 2: For IC value for both hyponyms and meronyms using the proposed IC formula,

$$\mathbf{IC(c) = 1- \frac{log(hypo(c)+mero(c)+1)}{log(max_{con})}}$$

go to step 3

Step 3: Call the existing semantic similarity function $Sim_{res}(c1,c2)$, $Sim_{Lin}(c1,c2)$, $Sim_{J\&C}(c1,c2)$

And then go to step 4.

Step 4: Call the proposed semantic similarity function for the given concepts c1 & c2

$$\mathbf{Sim_{ext}(c1,c2) = IC(LCS(c1,c2))-}$$

$$\sum_{hyper(c1,c2)}^{LCS(c1,c2)} \mathbf{hyper(c1) + hyper(c2)-}$$

$$\mathbf{hyper(c1 \cap c2)}$$

Step 5: Collect human judgments and save it as a separate table for the R&G and M&C data sets

Step 6: Calculate the correlation coefficients between results of the similarity measures and human judgments

Step 7: Compare the similarity measures for R&G data set using the proposed IC and proposed similarity existing similarity measures.

corpus may be present in alternate corpora like Reuters 20 news groups and google.

In WordNet taxonomy they do not take into consideration all the relation like meronymy.Since we are considering all the relationship present in the WordNet taxonomy, the new metric gives accurate results.

6. Conclusion

This paper has discussed the various approaches that could be used for finding similar concepts in an ontology and between ontologies. We have done a survey to exploit the similarity methods for ontology based query expansion to aid better retrieval effectiveness of Information retrieval models. The experiments conducted by early researches provide better correlation values which gives promising direction of using them in Ontology based retrieval models. A new semantic similarity metric has been introduced which overcomes the shortcomings of existing semantic similarity methods mainly sparse data problem. We are working with the new similarity function which will combine the advantages of the similarity methods discussed in this paper and we will test it with ontologies of particular domain.

7. REFERENCES

[1] Roy Rada, H. Mili, Ellen Bicknell, and M. Blettner. “Development and application of a metric on semantic nets” IEEE Transactions on Systems, Man, and Cybernetics, 19(1), 17{30}, January 1989.

[2] Michael Sussna: “Word sense disambiguation for tree-text indexing using a massive semantic network” In Bharat Bhargava

[3] Zhibiao Wu and Martha Palmer. “Verbs semantics and lexical selection”, In Proceedings of the 32nd annual meeting on Association for Computational Linguistics”, Association for Computational Linguistics, 1994. pages 133{138, Morristown, NJ, USA

[4] Claudia Leacock and Martin Chodorow: “Combining local context and wordnet similarity for word sense identification”. In Christiane Fellbaum, editor, WordNet: An Electronic Lexical Database. MIT Press, 1998.

[5] Philip Resnik.: “Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language”, 1999.

[6] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and K.J. Miller.[Resnik, 1995] Philip Resnik. “Using information content to evaluate semantic similarity in a taxonomy”. In IJCAI, pages 448{453, 1995.

[7] D. Conrath J. Jiang.: “Semantic similarity based on corpusstatistics and lexical taxonomy”. In Proceedings on

International Conference on Research in Computational Linguistics, Taiwan, pages 19{33, 1997.

[8] Dekang Lin.: “An information-theoretic definition of A similarity”. In Jude W. Shavlik, editor, ICML, pages 296{304. Morgan Kaufmann, 1998. ISBN 1-55860-556-8.

[9] Graeme Hirst and David St-Onge: “Lexical chains as representation of context for the detection and correction malapropisms” In Christiane Fellbaum, editor, WordNet: An Electronic Lexical Database. MIT Press, 1998.

[10] Troels Andreasen, Rasmus Knappe, and Henrik Bulskov: ” Domain-specifiic similarity and retrieval”. In Yingming Liu, Guoqing Chen, and Mingsheng Ying, editors, 11th International Fuzzy Systems Association World Congress, volume 1, pages 496{502, Beijing, China, 2005. IFSA 2005, Tsinghua University Press.

[11] Rubinstein and Goodenough, H. Rubinstein and J. B. a Goodenough: “Contextual correlates of synonymy”. Communications of the ACM, 8(10), 1965.

[12] George A. Miller and W. G. Charles: “Contextual correlates of semantic similarity”. Language and Cognitive Processes, 6(1):1{28, 1991.

[13] Dekang Lin: “Using syntactic dependency as local context to resolve word sense ambiguity”. In ACL, pages 64{71, 1997.

[14] Roy Rada and Ellen Bicknell: “Ranking documents with a thesaurus”.JASIS, 40(5):304{310,1989.Timothy Finin, and Yelena Yesha, editors, Proceedings of the 2nd International Conference on Information and Knowledge

a. Management, pages 67{74, New York, NY, USA, November 1993. ACM Press.

[15] Alexander Budanitsky: “University of Toronto Graeme Hirst, University of Toronto Evaluating WordNet-based Measus Of Lexical Semantic Relatedness”, Association for Computational Linguistics,2006

[16] Philip Resnik: “Using Information Content to Evaluate Semantic Similarity in a Taxonomy”, Sun Microsystems Laboratories, 1995

[17] Henrik Bulskov Styltsvig: “Ontology-based Information Retrieval Computer Science Section Roskilde University”, 1996

[18] A. Budanitsky. “Lexical semantic relatedness and its application in natural language processing”, 1999.

[19] A. Budanitsky. Semantic distance in wordnet: An experi mental, application-oriented evaluation of ¯ve measures, 2001.

[20] H. Bulskov: “Ontology-based Information Retrieval”, PhD Thesis. Andreasen, T., Bulskov, H., & Knappe, R. (2006).

[21] N. Seco, T. Veale, J. Hayes: “ An intrinsic information content metric for semantic similarity in WordNet, in : Proceedings of ECAI, 2004, pp. 1089–1090.

[22] Giuseppe Pirró: “A semantic similarity metric combining eatures and intrinsic information content”, Data & Knowledge Engineering 68(2009) 1289–1308 2009.