ISSN No. 0976-5697

Volume 3, No. 1, Jan-Feb 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A novel explicit context based approach for Computing Semantic Relatedness

Rezvan Mohamad Rezaei*
Department Of Computer Engineering
Science and Research Branch Islamic Azad University
Khozestan, Iran
r.mohamadrezaei@khouzestan.srbiau.ac.ir

Mehran Mohsenzadeh
Department Of Computer Engineering
Science and Research Branch Islamic Azad University
Tehran, Iran
mohsenzadeh@srbiau.ac.ir

Mashallah Abbasi Dezfouli
Department Of Computer Engineering
Science and Research Branch Islamic Azad University
Khozestan, Iran
Abbasi_masha@yahoo.com

Abstract: Computing semantic relatedness (SR) of words is a main functionality of large amounts of language applications. Explicit Semantic Analysis (ESA) is successful in computing semantic relatedness. ESA is an approach to explicitly represent the meaning of any text as a weighted vector of Wikipedia-based concepts and calculate semantic relatedness between terms or documents based on comparing the corresponding vectors. However, ESA method generates the same vector for an ambiguous word and does not consider the given context of the word-pair. In this paper, we propose an improved method that first specifies the given context of word-pairs by Wikipedia-based concept network named wikinet then represents any word as a weighted vector of Wikipedia-based concepts and compares vectors by cosine metric. Empirical Evaluation on WordSimilarity-353 dataset shows that the proposed method provides consistent improvements in correlation of computed relatedness scores with human judgments compared to the existing methods.

Keywords: computing semantic relatedness, wikinet, Wikipedia, word sense disambiguation, explicit semantic analysis.

I. INTRODUCTION

All The need for determining semantic relatedness between two words is a problem that pervades much of natural language processing such applications as word sense disambiguation, determining the structure of texts, text annotation, information extraction and retrieval, automatic indexing, lexical selection, and the automatic correction of word errors in text [14].

Computing semantic relatedness of (jaguar, cat) or (jaguar, car) is ordinarily performed by humans due to their knowledge and experience. To perform it by Computers, they require access to large amounts of world knowledge. Previous approaches to computing semantic relatedness used various analysis methods. Some of approaches use the knowledge based on statistical analysis of lare corpora [1], [2]. LSA does not rely on any human-organized knowledge. There is another approach based on handcrafted lexical resources such as taxonomies and thesauri that incorporates very limited knowledge about the world [3], [4]. The drawback of this approach is that creation of lexical resources requires lexicographic expertise and effort, and consequently such resources cover only a small fragment of the language lexicon [5].

Some of approaches represent meaning in a space of natural concepts derived from Wikipedia[i] [5], [6], and [7].

The Explicit Semantic Analysis (ESA) [5] is a promising approach for calculating semantic relatedness. Due to the use of natural concepts, ESA is easy to explain to human users. Nevertheless, ESA does not consider the given context of the word-pair and generates the same vector for an ambiguous word in different word-pairs.

Self-Adapting Explicit Semantic Analysis (SAESA) [8] generates corresponding concepts for one word in different context, according to the different words being compared. It achieves a higher accuracy than the ESA method. However, SAESA method must consider all of the concepts of Wikipedia twice. So it consumes high cost and time. ESA and SAESA methods only make use of the article information that Wikipedia contains. In this paper, in addition to the articles we used the categorization structure of the articles. Therefore, we introduce an improved approach that firstly specifies the given context of word-pair using the concepts and their relations are extracted from Wikinet. Wikinet is a multi-lingual concept network obtained automatically by mining for concepts and relations and exploiting a variety of sources of knowledge from Wikipedia [10]. Then based on specified context of word-pair, it interprets each word of word-pair as weighted vector of Wikipedia based concepts and compares corresponding vectors using the cosine metric.

The remainder of this paper is organized as follows. Section 2 firstly provides background on ESA and wikinet and then surveys previous methods. Sections 3 and 4 describe the details of the proposed algorithm and the evaluation results. Finally, Section 5 concludes the paper.

II. RELATED WORK

A. Explicit Semantic Analysis:

Explicit Semantic Analysis, or ESA [5], is a recently proposed method for computing the semantic relatedness. ESA represents meaning in a high-dimensional space of concepts, automatically derived from large-scale human-built repositories such as Wikipedia. ESA is based on the assumption that in Wikipedia an article corresponds to a semantically distinct concept. Each article is pre-processed by tokenization, stemming and stop word removal and is represented as a vector of words that occurs in this article weighted by their TFIDF score. These weights quantify the strength of association between words and concepts.

Firstly, a text fragment is represented as a vector using TFIDF scheme, then semantic interpreter that implements as a centroid based classifier [9] ranks all the Wikipedia articles by their relevance to the text. To speed up semantic interpretation, ESA uses an inverted index, which maps each word into a list of concepts in which it appears. The semantic interpreter iterates over the text words, retrieves corresponding entries from the inverted index, and merges them into a weighted vector of concepts that represents the given text. Entries of this vector reflect the relevance of the corresponding articles to text. This method compares weighted vectors of the Wikipedia articles related to a particular term or portion of text using the cosine metric to compute semantic relatedness. Gabrilovich and Markovitch show that ESA outperforms other existing approaches.

Compared with the previous state of the art, using ESA results in notable improvements in correlation of computed relatedness scores with human judgments. However, ESA doesn't specify a given semantic context of the word-pairs and uses the similar vectors for ambiguous word.

Consider, for example word-pairs such as (jaguar, cat) and (jaguar, car). The semantic context of (jaguar, cat) is animal and the semantic context of (jaguar, car) is automobile. However, ESA when computing semantic relatedness builds similar vectors for "jaguar" without paying any attention to the semantic context of words. In this paper we introduce an approach that firstly specifies the given context of word-pair using the Wikinet.

Wikinet:

In recent years, researchers have realized the huge potential of Wikipedia as a source of semi-structured knowledge and several systems have used it as their main source of knowledge. Wikipedia provides a massive and relatively high-quality collection of text and encyclopedic knowledge. The use of a knowledge repository as large and diverse as Wikipedia creates a powerful concept network, well suited for computing semantic analysis.

First, Wikipedia's broad coverage of a huge range of topics, and second, mapping from a massive aggregation of natural language terms to the concepts in which they occur, produce a powerful classifier to automatically map any text fragment to this concept network. Finally, the use of a Wikipedia generates meaningful and human readable concepts that can provide additional reasoning for the researcher and for system users.

Several approaches have been used to extract semantic information from Wikipedia. Wikinet [10] is a very large scale, multilingual concept network, obtained by exploiting several facets of Wikipedia. The resource consists of a language independent concept base extracted from Wikipedia articles and categories, and the relations between them. Relations between concepts are extracted from the several sources of knowledge from Wikipedia some explicitly (articles, categories and their links, infoboxes), some implicitly (category names). This coverage of multiple pieces of information differentiates it from similar endeavors and resources extracted from Wikipedia. WikiNet is supposed to be used to complement WordNet with knowledge about numerous named entities (which were outside the scope of WordNet) as well as general concepts and numerous relations.

In this paper, we made use of wikinet for specifying given semantic context of word-pair because Wikipedia is only partly structured and cannot be used by a computer without some processing. This concept network provides concepts and relations that related to any word of word-pair. Then we compare concepts and their relations and generate a given context of word-pair.

Previous Methods:

Use Automatically estimating semantic relatedness has been fundamental problem for decades, and has been addressed by diverse techniques in cognitive science, computational linguistics, artificial intelligence, and information retrieval. For example, in computational linguistics, applications of sematic relatedness include word sense disambiguation, information retrieval, word and text clustering [14]. Until recently, computing semantic relatedness of natural language words required encoding large amounts of common sense and domain specific world knowledge.

Prior work pursued three main directions: using statistical analysis of large corpora to provide the knowledge base [1] such as using Latent Semantic Analysis (LSA) [2], using hand-crafted lexical resources such as WordNet [14], and using Wikipedia as knowledge base [5], [6], [7]. LSA is a purely statistical technique, which leverages word co-occurrence information from a large unlabeled corpus of text. LSA does not rely on any human organized knowledge; rather, it "learns" its representation by applying Singular Value Decomposition (SVD) to the words-by-documents co-occurrence matrix [2]. Latent semantic models are difficult to interpret, since the computed concepts cannot be readily mapped into natural concepts manipulated by humans.

Some of approaches are using hand-crafted lexical resources such as taxonomies and thesauri [4] [14]. Some ways to compute semantic relatedness in taxonomy such as WordNet are [14]: 1) view it as a graph and identifying relatedness with path length between the concepts 2) Considering the extent of shared information in common 3) view it as a network and scaling the links in the taxonomy. The obvious drawback of these approaches is that creation of lexical resources requires lexicographic expertise as well as a lot of time and effort, and consequently such resources cover only a small fragment of the language lexicon. Specifically, such resources contain few proper names, neologisms, slang, and domain-specific technical terms. Furthermore, these resources have strong lexical orientation and mainly contain information about individual words but little world knowledge in general [5].

These limitations are the motivations behind several new techniques [5], [6], [7] that use Wikipedia, the largest encyclopedia in existence, as the knowledge base. Authors of [7] explore the idea of using Wikipedia for computing semantic relatedness. Semantic relatedness is then computed using various measures that either relies on the texts of the articles that contain words in their titles, or path distances within the category hierarchy of Wikipedia. WikiRelate! Is limited to single words and represents the semantics of a word by either the text of the article associated with it, or by the node in the category hierarchy. [6] Uses Wikipedia to provide a large amount of structured world knowledge about the terms of interest and is the first procedure that is using only the hyperlink structure of Wikipedia rather than its full textual Content.

The Explicit Semantic Analysis (ESA) [5] described in section A. in this method, Due to the use of natural concepts, it is easy to explain to human users. Nevertheless, ESA does not consider the given context of the word-pair and generates the same vector for an ambiguous word in different word-pairs. This limitation is the motivation behind Self-Adapting Explicit Semantic Analysis (SAESA) [8] that generates corresponding concepts for one word in different context, according to the different words being compared. SAESA method has a higher accuracy than the ESA method and is more suitable for ambiguous words. However, SAESA method must consider all of the concepts of Wikipedia twice. So it consumes high cost and time. ESA and SAESA methods only make use of the article information that Wikipedia contains, i.e. only analyzes the term→article allocation. However, Wikipedia provides a wealth of semantic information, namely the links between articles and the categorization structure of articles.

III. PROPOSED METHOD

First As described above, Explicit Semantic Analysis (ESA) has got a great success in computing the semantic relatedness based on Wikipedia but ESA neglects the context of word-pair, thus it cannot exactly determine the desired sense of an ambiguous word. Therefore, we introduce our method that first specifies the context of words semantically, then extracts corresponding concepts for each Word under the given context and computes Semantic relatedness of word-pair. We use the wikinet for Specifying given context of words. The detail of the method is as follows.

Step1: Word-pair Category Specification

First, The Initial step of our method is Word-pair Context Category. The 'Word-pairCategorySpecification' algorithm chooses a given category for related concepts by extracting concepts and their relations from WikiNet. The algorithm firstly finds indexes of words in word-pair from index.wiki file and saves in $Index_{w1}$, $Index_{w2}$. If a word of word-pair has several indexes, it is an Ambiguous word [10], therefore we must specify intended sense of the word. Then it extracts relations indexes from data.wiki file and each concept of $Index_{w1}$ and $Index_{w2}$ compared based on IS-A relations. If two $concept_1$ and $concept_2$ have IS-A relation with a given concept, they are related concepts and have relation on a given context [10], [15], [16]. Although the IS-A relation to a common concept does not exist between some of concepts, they are somehow related. Thus, after considering above condition, we extract from inlinks.wiki file the list of concepts that link to $concept_1$ and $concept_2$ through their corresponding pages and save in $InLink_{concept1}$ and $InLink_{concept2}$. Now it compares the IS-A relations of $InLink_{concept1}$ to IS-A relations of the $concept_2$ and vice versa. The given concepts that have IS-A relation to both words indicates the given context for words of word-pair [15], [16].

$InLink_{concept1}$ **R** $Concept_1$

$InLink_{concept1}$ **IS-A** X

$Concept_2$ **IS-A** X

$\Rightarrow$ $InLink_{concept1}$ **R** $Concept_2$

$InLink_{concept1}$ **R** $Concept_1$

$InLink_{concept1}$ **R** $Concept_2$

$\Rightarrow$ $Concept_1$ **R** $Concept_2$

All Finally, we choose a given category for related concepts that the category's articles have relation with related concepts. We define C_1 and C_2 as the set of categories assigned to $concept_1$ and $concept_2$, respectively. We then determine the semantic relatedness value for each category pair (c_k, c_l) with $c_k \in C_1$ and $cl \in C_2$. Then we uses the notion of a lowest common subsumer of two nodes lcs (c_k, c_l).i.e. the minimum for path based [11]. The given category is a common semantic environment for related concepts.

Algorithm : Word-pairCategorySpecification (w_1,w_2)

For each $index_{w1i} \in Index_{w1} = \{ index_{w11},\ldots,index_{w1n} \}$ do
  For each $index_{w2j} \in Index_{w2} = \{ index_{w21},\ldots,index_{w2m} \}$ do
    Let $Is\text{-}aIndex_{w1i}$ be a set of concepts that have is-a relation with $index_{w1i}$
    Let $Is\text{-}aIndex_{w2j}$ be a set of concepts that have is-a relation with $index_{w2j}$
    $CommonConcept_{indexw1i,indexw2j} \leftarrow Is\text{-}aIndex_{w1i} \cap Is\text{-}aIndex_{w2j}$
    If $CommonConcept_{indexw1i,indexw2j} \neq \emptyset$ then
    Let $InLink_1$ be a set of concepts that link to $index_{w1i}$
    Let $InLink_2$ be a set of concepts that link to $index_{w2j}$
    For each $InLink_{1x} \in InLink_1 = \{InLink_{11},\ldots, InLink_{1s}\}$
      Let Is-a$InLink_{1x}$ be a set of concepts that have is-a relation with $InLink_{1x}$
      $CommonConcept_{indexw1i,indexw2j} \leftarrow Is\text{-}aInLink_{1x} \cap Is\text{-}aIndex_{w2j}$
    For each $InLink_{2y} \in InLink_2 = \{InLink_{21},\ldots, InLink_{2t}\}$
      Let Is-a$InLink_{2y}$ be a set of concepts that have is-a relation with $InLink_{2y}$
      $CommonConcept_{indexw1i,indexw2j} \leftarrow Is\text{-}aInLink_{2y} \cap Is\text{-}aIndex_{w1i}$
    If $CommonConcept_{indexw1i,indexw2j} \neq \emptyset$ then
      GivenCategory ← CommonCategory(categories of $indexw_{1i}$,$indexw_{2j}$)
      $Commonconcept \leftarrow CommonConcept_{indexw1i,indexw2j}$
If $Commonconcept = \emptyset$ then
GivenCategory ← CommonCategory(categories of $Index_{w1}$, $Index_{w2}$)
Return GivenCategory

Step2: Word-pair Context Word Generation

The 'Word-pair Context Word Generation' algorithm specifies context words based on the given category. In this algorithm, we measure the cosine metric of the TFIDF vectors of the word-pair and the articles of given category to determine their relatedness. Finally, in order to utilize the context information in the Context Related Article, we select a set of B words with the highest TFIDF weight in Context Related Article to represent the context of the word-pair.

Similar to [8], In proposed method, we use standard attribute selection techniques such as information gain to identify words that are most characteristic of a concept versus of all other concepts.

Algorithm : Word-pairContextWordGeneration (w_1,w_2)

GivenCategory ← Word-pairCategorySpecification (w_1,w_2)
Text(w) ← Combine(Text(w_1) , Text(w_2))
Vector(w) ←TFIDF(Text(w))
For each $concept_i \in$ GivenCategory = {$concept_1$,…, $concept_n$} **do**
  Text($concept_i$) ← AttributeSelection(Text($concept_i$))
  Vector($concept_i$) ← TFIDF(Text($concept_i$))
  Value($concept_i$) ← DistanceMetric(Vector($concept_i$), Vector(w))
Let *ContextRelatedArticle* be a set of A articles that have highest value($concept_i$)
For each $concept_i \in$ *ContextRelatedArticle* **do**
  Text(ContextRelatedArticle) ← add(Text($concept_i$))
  Vector(ContextRelatedArticle) ← TFIDF(Text(ContextRelatedArticle))
Let *ContextRelatedtext* be a set of B words that have highest TFIDF in the Text(ContextRelatedArticle)
Return *ContextRelatedtext*

Step 3: Word Concepts Generation

After acquiring the ContextRelatedtext by Step 2, we must generate the corresponding concepts for each word of word-pair with the help of the ContextRelatedtext that generated semantic context of the word-pair and GivenCategory. The 'WordConceptsGeneration' algorithm generates the corresponding concepts for each word of word-pair.

Algorithm: WordConceptsGeneration

ContextRelatedtext ← Word-pairContextWordGeneration (w1,w2)
Text(w_1) ← combine(Text(w_1), *ContextRelatedtext*)
Text(w_2) ← combine(Text(w_2), *ContextRelatedtext*)
Vector(w_1) ← TFIDF(Text(w_1))
Vector(w_2) ← TFIDF(Text(w_2))
For each $concept_i \in$ GivenCategory = {$concept_1$,…, $concept_n$} **do**
  Text($concept_i$) ← AttributeSelection(Text($concept_i$))
  Vector($concept_i$) ← TFIDF(Text($concept_i$))
  $Value_{W1}$($concept_i$) ← DistanceMetric(Vector($concept_i$),Vector(w_1))
  $Value_{W2}$($concept_i$) ← DistanceMetric(Vector($concept_i$), Vector(w_2))
Let $Vector_{w1}$ = {$Value_{W1}$($concept_1$),…., $Value_{W1}$($concept_n$)}
Let $Vector_{w2}$ = {$Value_{W2}$($concept_1$),…., $Value_{W2}$($concept_n$)}
Return $Vector_{w1}$, $Vector_{w2}$

Step 4: Computing Semantic Relatedness

In "The Computing Semantic Relatedness" step, in order to compute semantic relatedness of a pair of words, we compare their vectors using the cosine metric.

SR = Cosine ($Vector_{w1}$, $Vector_{w2}$).

To illustrate it, we show the common concepts and the concepts of ambiguous word vector for two sample Word-pairs: (jaguar, tiger) and (jaguar, car). Table I shows the common concepts that are extracted from IS-A relations. These word-pairs contain ambiguous word "jaguar". In our method, the semantic environment of word-pair specified by information of a common context therefore it is capable of performing word sense disambiguation. Table II contains the ten highest-valuing Wikipedia concepts in $Vector_{w1}$ for "jaguar". If concepts in the vector are sorted in the decreasing order of their value, the top ten concepts are the most relevant ones for the word of word-pair.

Table: 1 Common Concepts that Extracted From is-A Relations for Sample Word-Pair

(jaguar, tiger)	*(jaguar, Aston Martin)*
animal	Car manufacturers
mammal	Automotive companies
fauna	Vehicle
biota	Road transport
Big cat	Motor vehicle manufacturers

Table: 2 The ten Highest-Valuing Wikipedia Concepts in $Vector_{w1}$ for "Jaguar"

(jaguar, tiger)	***(jaguar, Aston Martin)***
Jaguar	Jaguar car
cougar	Jaguar S-Type
leopard	Jaguar E-Type
genus Panthera	Jaguar X-Type
Big cat	Jaguar C-Type
Snow leopard	Jaguar XK
Panther hybrid	Jaguar XJ
tiger	Jaguar XF
puma	V8 (V8 engine)
European lion	Luxury vehicle

Figure 1 makes clear our approach and illustrates the process of computing semantic relatedness of word-pairs by using wikinet and Wikipedia.

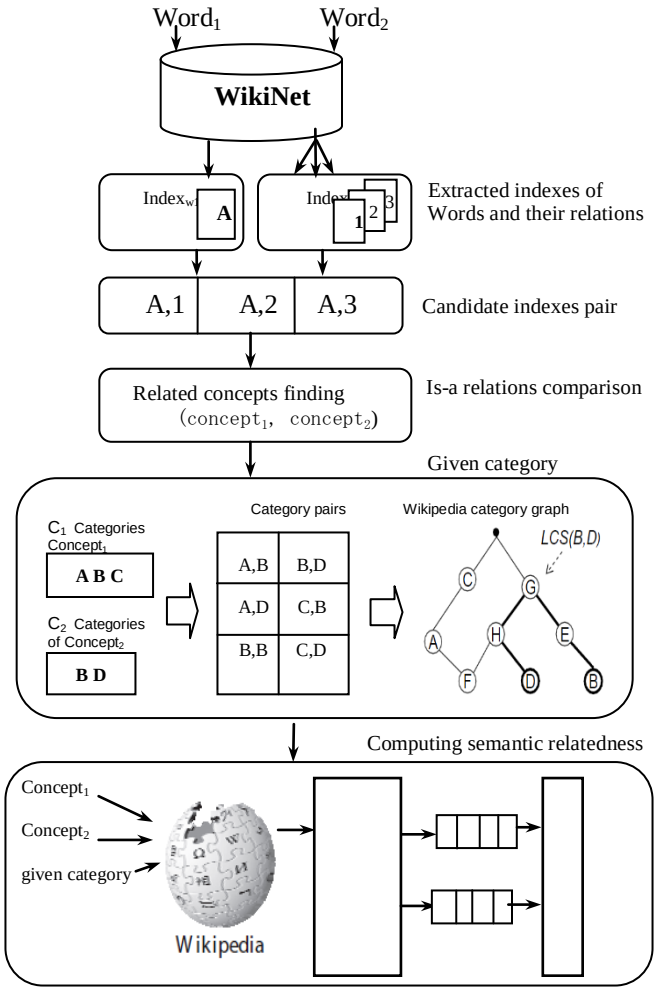

Figure 1. the process of method

IV. EMPIRICAL EVALUATION

In this paper, we evaluated the method for the application of computing semantic relatedness between word-pairs. For implementation of our approach, we use a Wikipedia [ii] snapshot as of "January 15, 2011" and "20110115" version of WikiNet [iii] . With WikiNet's 3.7 Million concepts and 40 Million relations (instantiating 656 relation types), efficiency in data management becomes an issue. Manual analysis of the data is also problematic. WikiNetTK[iv] [13] addresses both these issues. A fast data management is the basis for an easy-to-use visualization component. For using wikipedia first, we preprocess the Wikipedia dump with stemming, stop word removal, frequency of words calculation and attribute vector discovering. So Wikiprep [v] parses the Wikipedia XML dump. Upon removing small and overly specific concept (those having fewer than 100 words and fewer than 5 incoming or outgoing links), In order to speed up the experiment process, we used the Natural Language Toolkit to parse each article as single words and their frequency of seeming in this page. We processed the text of these articles by first tokenizing it, removing stop words and rare words (occurring in fewer than 3 articles), and stemmed the remaining words, this giving up distinct terms, which were used to represent article as attribute vectors.Additionally, since the dataset is large enough to be portioned into training set and testing set, the pattern selection for optional parameter estimation was performed as a grid search through cross-validation on the training data. In experiment, the value of A is 9 and the value of B is 16.

Dataset And Evaluation Procedure

Evaluating word relatedness is a natural ability humans have and is, therefore, considered a common baseline. To assess word relatedness, we use the WordSimilarity-353 benchmark dataset, available online [12], which contains 353 word pairs. Each pair was judged, on average, by 13-16 human annotators. This dataset, to the best of our knowledge, is the largest publicly available collection of this kind. Spearman rank-order correlation coefficient between the computed relatedness scores by experiment and the corresponding human judgments on WordSimilarity-353 benchmark were used to compare computed relatedness scores with human judgments.

Result

Table III shows the result of applying our method for making judgment relatedness of word-pairs of WordSimilarity-353.

Table: 3 Result onCorrelation with Human Judgements of Relatedness Measure

algorithm	Correlation with humans
WordNet [4]	0.33-0.35
Roget's Thesaurus [4]	0.55
LSA [17]	0.56
WikiRelate! [7]	0.19-0.48
WLVM [6]	0.45
ESA-wikipedia [5]	0.75
SAESA-wikipedia [8]	0.81
Our method	0.81

The results of experiment show that our method makes substantial improvements over prior studies. Our method, compared to statistically methods such as LSA, which only uses statistical co-occurrence information from a large unlabeled corpus of text, uses the knowledge resource that is collected and organized by humans. Also, compared to the methods relying on lexical resources such as WordNet and Roget's Thesaurus that cover only a small fragment of the language lexicon, our method leverages knowledge resources that are orders of magnitude larger and more comprehensive. Our proposed approach improved also Wikipedia-based approach such as WikiRelated! [7] and WLVM [6]. That is because our approach represents intended sense of each word as a weighted vector of Wikipedia concepts, and semantic relatedness is then computed by comparing the two concept vectors. Besides, the obtained results of ESA method is most close to our method and shows that our method is more effective to computing semantic relatedness. Our method also achieves similar results to that of the SAESA. However, SAESA in order to generate given context of word-pair considers all concepts in Wikipedia that it needs for a very long time. But our method finds the given context of word-pairs just by searching the information of wikinet, and based on the context, it expresses the intended meaning for the ambiguous word and compute semantic relatedness. However, Empirical evaluation confirms that using our method leads to substantial improvements in computing semantic relatedness of word-pairs.

V. CONCLUSION

In this paper, we proposed an improved approach to compute semantic relatedness of words using information of Wikipedia. This approach before extracting corresponding concepts for each word, by using of a Wikipedia-based concept network named wikinet, generates the given context of word-pair and specifies the intended sense of ambiguous word. ESA generates the same semantic interpretation vector for ambiguous words, but our proposed method has made a great improvement in relation to computed relatedness scores with human judgments, from r = 0.75 to 0.81. We achieve same results as SAESA, but this approach does not consider all of the articles in Wikipedia and just considers the relations of concepts. In this paper, in addition to the articles we used the categorization structure of the articles to specification a given semantic environment (context) of word-pairs.

VI. REFERENCES

[1]. R. Baeza-Yates, B. Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley, New York, NY, 1999.

[2]. S. Deerwester, S. Dumais, G. Furnas,T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," JASIS, 41(6):391–407, 1990.

[3]. C.Fellbaum, "WordNet: An Electronic Lexical Database", MIT Press, Cambridge, MA, 1998.

[4]. M. Jarmasz."Roget's thesaurus as a lexical resource for natural language processing," Master's thesis, University of Ottawa, 2003.
[5]. E. Gabrilovich, S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis," IJCAI, pp.1606-1611,2007.
[6]. D. Milne, "Computing semantic relatedness using Wikipedia link structure".NZCSRSC'07, 2007.
[7]. M.Strube, S.P.Ponzetto, "WikiRelate! Computing semantic relatedness using Wikipedia," In AAA'06, Boston, MA, 2006.
[8]. W.Wang, P.Chen, B.Liu, "A Self-Adaptive Explicit Semantic Analysis Method for Computing Semantic Relatedness using Wikipedia" International Seminar on Future Information Technology and Management Engineering, 2008.
[9]. Eui-Hong (Sam) Han, G.Karypis,"Centroid-based document classification: Analysis and experimental results,"In PKDD'00, September 2000.
[10]. V.Nastase, M.Strube, B.B¨orschinger, C.Zirn, A. Elghafari ,"WikiNet: A Very Large Scale Multi-Lingual Concept Network," In Proceedings of the 7th International Conference on Language Resources and Evaluation, La Valetta, Malta, 2010.
[11]. T. Zesch, I. Gurevych."Analysis of the Wikipedia Category Graph for NLP Applications". In Proc. of the TextGraphs-2 Workshop, NAACL-HLT,2007.
[12]. L.Finkelstein, E.Gabrilovich, Y.Matias, E.Rivlin, Z.Solan, G.Wolfman, E.Ruppin, "Placing search in context: The concept revisited". ACMTOIS, 20:116-131, 2002.
[13]. A.Judea, V.Nastase, M.Strube, "WikiNetTK- A ToolKit for Embedding World Knowledge in NLP Applications,"Proceedings of the IJCNLP 2011 System Demonstrations, pages 1–4,Chiang Mai, Thailand, November 9, 2011.
[14]. A. Budanitsky, G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," Computational Linguistics, 32(1):13–47, 2006.
[15]. O. Medelyan, D. Milne, C. Legg and I. H. Witten,"Mining meaning from Wikipedia" International Journal of Human-Computer Studies. Volume 67, Issue 9, pp. 716-754, 2009.
[16]. S.Paolo Ponzetto, M.Strube,"knowledge derived from Wikipedia for computing semantic relatedness," Journal of Artificial Intelligence Research, 181-212, 2007.
[17]. H. Kozima, and T. Furugori, "Similarity between words computed by spreading activations on an English dictionary", EACL-93, 1993, pp.232-239.

