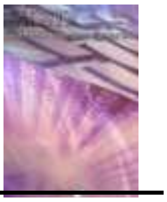
DSR Design for Scientific RENAISSANCE



A NEW METHOD FOR THE CONSTRUCTION OF RELATEDNESS SEMANTIC NET. BASED ON DISTRIBUTIONAL SEMANTICS

Bahaa A. Mohammed, Ahmed T. Sadiq, Mohammed Natiq

Department of Computer Science, University of Technology-*Iraq*

ISSN (Printed): 2314-7350
ISSN (Online): 2231-8852

Abstract

In this paper, proposal of local semantic net. created by a new special algorithm called automatic semantic net algorithm (ASN). finding the relevance degree between words in self-document without background knowledge is the focus of attention of this proposal, depends mainly on the characteristics of distributional semantic which contributed to build a specific word-net for each document. In addition to that taking into account the convergence between the sequence of words in document addition to frequency of each word. used the amendment on (Wu & Palmer) similarity metric measurer to make it commensurate with the proposed semantic net. experimental results it was very convincing compared to prior knowledge semantic net approach and the modify on standard similarity metric measure gave better results.

Keywords: distributional semantic, semantic net, semantic measure, semantic relatedness.

1. Introduction

Mostly all information and features that important to us its finding in text. therefore, the text analysis takes a large area in recent re- searches about data mining and machine learning to processing a huge amount of information and extraction featured of which (Xiaofei, 2014). Semantic approach is one of main processes in this field, it's give a great addition to text classification and clustering, by extract the relation and the relevance degree among different words depending on implicit meaning of these words. a lot of problems facing this operation when it comes to the meaning of word. Because all the natural languages contain a many ways to express the meaning by different of phrase and words (Tao, 2004).

In general the semantic similarity approach divided in two main category : the first, its contain a priori knowledge for each word or topic that most common by assign background concept to the words that represent a public notion in special topics, and this understanding representation be significant to text analysis such as clustering and classification of the texts, for example the word "machine" when see this word initiate to our minds many other words involving under concept of this word "machine" it was represent public idea needed to clarify. The second category of semantic it's without prior knowledge, it's based on obtaining type of understanding statistically based on corpus, corpus is a one of main approaches to measure the similarity between terms according to large sets of text (Wael, 2012; Wael, 2013), the distributional semantic which this proposal based on it belongs to corpus's family it's based on geographical distribution of the words in the text to determine the semantic measure between pair words, this approach depending on distributional hypothesis which said the understanding of word presence in set of other words it which occurs in same context with this word (Marco, 2010; Peter, 2010), excellent results appeared this approach in recent researches by combined distributional semantic with latent semantic analysis (LSA) and probabilistic variant LSA specially in web mining, its used terms that frequents together almost to extract automatic concept about these terms from documents croups by (PLSA) (Bamshad, 2004).

At start created a semantic net to representing concept of document's words (Simmons , 1978), using the proposed algorithm (ASN) based on geographic distribution of words in the text and compute semantic measures using modify on (Wu & Palmer) measurement method (Wu, 1994). in the result will get relatedness degree of each pair word which represents semantic degree between the words that belongs to same text, better results compared to standard knowledge-based semantic net (WordNet, 2016) with standard similarity measurer (Wu & Palmer) (Wu, 1994), to obtaining the semantic ratio.

2. Related works

There are several semantic approaches gained the great concern in the recent researches that interested in text analysis to importance of this domain in support of data mining and machine learning. will discussing some this approaches in this section.

In (Tao, 2010), is a Local Latent Semantic Indexing (Local LSI) by Tao Liu Zheng hen and others, proposed a modify on standard latent semantic indexing (LSI) to enhance the classification of documents by applied performing a separate Single Value Decomposition (SVD). in the initial of local LSI, classified each documents from data set to the relevant topic, and build a local classified region such as (sports, political, society, health and so on), the first divided was gave a clearer conception to Support categorization of text. The proposal results improved the execution time of LSI methods. term vector is the fastest and it needs only hundred seconds. Global LSI needs much more time than term vector due to the costly SVD computation on entire training set. Although SVD computation on local region is very fast, the overall computation on all topics is extremely high.

In (Evgeniy, 2006), its Explicit Semantic Analysis (ESA), this type of semantic build a semantic interpreter by create a fragment maps of natural language text in other word its build a net for each word and connected other words by relevant in field of used or in meaning or depended on order concept, after fragment text represent as weighted vectors of concept called (interpreter vectors) and create special group for each word depend on one of The foundations mentioned above and computing semantic relatedness by amount of vectors in space that defined by concept, in other understanding this method takes two words and measured the semantic degree between them by compute number the related words in the net of each one.

In (Peng, 2015), they are prepared a solution for Short texts, usually the short text faced a problem in finding the semantic of word that relevant with this short phrases, to solve this issue by using approach called embedding space, its mean adds a predicate context for short text depend on semantic between them. For example:

$$vec(Germany) + vec(Capital) \approx vec(Berlin).$$

In above example the word *berlin* represents an embedded space of this short sentence, by using clustering algorithm CNN to embedding and discover semantic cliques.

In field of learning semantic models (Paola, 2011), prepared a semantic learning by build a specific or local semantic for each input text and depend on same text to compute semantic between word by create local or self-WordNet and used as source for relevant word in the text domain, and this method gives a pure concept for each field and its appear More accurate results because its dealing with each input by special semantic Commensurate with each field, Wikipedia has been used to enrich text representations with concepts to improve supervised text.

In (Xinghua , 2015), they prepared text categorization with semantic enriched, this method of semantic based on Latent Dirichlet Allocation (LDA), it's semantic approach used a small numbers of topics to represent a collection of documents by take a multi-topical features of documents and treats a document as a mixture of words, in LDA concept represent a large number of documents by one topic and this done when take a two biggest words probability from all of documents and find a topic can include these biggest two word in one filed. And Amal et al. (Amal, 2017), present a proposal based on neural network to improve the bag-of-word approach and enhanced the embedding words to make it updates depended on the time, for example the word "trump" before ten years Was associated with money and business, but today it is associated with politics.

3. Distributional Semantic Model (DSM)

This type of semantic approach rely on the distributional hypothesis to computing the relatedness measure between words, the distributional hypothesis assumes the meaning or concept of word presence in implicit words of the text in similar context of the sentences, the DSM represented by special matrix, rows of this matrix is contain words as targets which collected by take sample of window with fixed size such as (three words) and passed on the text's words, and the words in columns represented the meaning or understanding of target words in the rows, by counting the number of co-occurrence between these words and convert the words to vector to compute the semantic between pair words (Douwe, 2014). To understanding the distributional structure will must consider the following: First, the words that reflect what in our thinking do not meet together in arbitrary: that's each term occurs in same positions in the relative to same others terms. Second, but also not ignore some terms that appear in it seems as arbitrarily, such as it was considered that the precise distribution is very difficult or impossible in the analysis of natural language because of some ambiguity that comes with the existence of some of terms in con- text. But this is not accurate. since the appearance of elements in the language must have a certain basis or reason based on the emergence of other terms. Third, it is possible to obtain a good degree of accuracy related to the occurrence of any element related to the occurrence of another element so that don't need to external sources of information to support the distribution structure of the language (Zellig , 1954).

Semantic net is an knowledge base to describe the concept of terms, it's a special technique to finding the relation between words by two form either mathematically or based on prior knowledge, semantic nets Includes hierarchal representation(fig.1) consist edges and nods to clarify the terms that belongs to similar understanding or it's have big presence with specific other terms, in semantic net model the nodes are presence in beginning of the net be more general to describe many terms, while the nodes in depth of net be more specific, So this relationship in semantic net its built either on the basis of one meaning or the convergence in the same concept, in this model attempted to create semantic net without a background knowledge, only through using the words of same text in the testing to find specific relation between words supported the distributional semantic approach (Patrick, 1992).

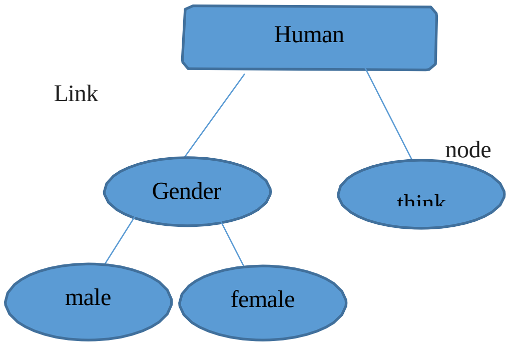


Fig. 1: hierarchal representation of concept.

The chosen method its Wu & Palmer (Wu, 1994) as similarity metric based on Rada measure, which depend on hierarchical structure to compute the semantic degree between two words by calculate the path between these two words, Wu & Palmer or (WUP) it's an efficient measurement method in hierarchical environment (fig.1), moreover it has a simple understanding and smooth implementation, in addition its gave suitable results with this proposal, for these reasons preferred this method from others similarity measures. following equation (1) shown the forma of this method:

$$sim\ wp(X,Y) = \frac{2(P)}{A+B} \qquad (1)$$

Where *sim-wp* is Wu & Palmer similarity measure, and *A*, *B* represent the path length from each concept *X*, *Y* to the root, where P it's a path length from common node between the two concept *X*, *Y* to the root LCS. (Canada, 2016), (fig.2) will illustrate the parameters of method.

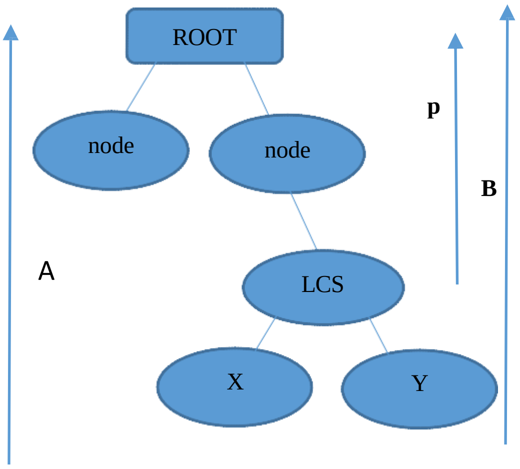


Fig. 2: Wu & Palmer parameters.

4. Proposed Method

This section discusses two main aspects of the proposal, will be explained in detail in the following:

4.1 Construction of Relatedness Semantic Net

This aspect from proposed illustrates act to build a special word- net represented the relations between words which belongs to one document, the essential objective its obtains a maximum relevance of semantic between words As well as weight and importance of the word relative to the text to which it belongs. This method is mainly based on related and sequential of speech in certain subjects basis of the languages it's a translation of human ideas to form of those texts and the speech must be coherent and consistent with the idea and subject matter of the text has been clarify the steps of proposed in the following points:

1. *Input the text after preprocessing and compute frequency of each word in text.*
2. *Get the maximum word frequency from text and as- sign it as a Root of the word-net.*
3. *Create a slid window with size only three words and passed from start to end of text.*
4. *For the first slid that contain three sequential words compute frequency for these words and get the max word as parent to others two words, and assign it as first node in word-net from the Root, thus been set up the first path in semantic net.*
5. *For other two words in this slid as well take the max word as parent to mini word based on occurrence of these words in the text, at the result obtained a node represented by max word in slid and others two children represented of max child as direct connected node by max parent and mini chilled as direct connected node by max chilled.*

6. *The slid is passed from max word in Which precedes it by one word.*
7. *if words appearance with same frequency in same slid then all words belonging to this slid assign node to each of one from last largest node.*
8. *If words of slid it's have a node found in this word- net then ignoring added to the word-net, but will be taking a value of these words in account to the next step.*
9. *this step has a three cases:*
10. *if a max parent of present slid is largest than all nodes in same path of last slid then create a new node for this present slid from Root with new path. And apply step 3.B for its children.*
11. *if max parent of present slid smaller than max parent in last slid and lager than others children then cre- ate new node from max parent of last slid to max present. and apply step 3.B for children.*
12. *if max parent of present slid smallest than all nodes in same path of last slid then connected to smallest node that larger from it and apply step 3 to children.*

Used the modify on (WUP) similarity metric method to compute the relevance degree between words.

4.2 semantic relatedness Measurement

The adopt on distributional semantic in this proposal to make an adjustment on the similarity measure which used in this model so that make it fit into the co-occurrence issue because it is one of ways to represent the distributional semantic through taking in account the terms that frequency together in the text, for the men- tioned above, then added a simple extension to solve this issue, as shown in the following equation (2):

$$\text{sim mwp}(x, y) = \frac{2(cp)}{A + B} + \frac{2 * f(xy)}{Fx + Fy} \qquad (2)$$

where *sim-mwp* is modify on the original wp, and the *CP* represent the number of common nodes between *X, Y* including the ROOT node, and the F*(XY)* is frequency of pair words *X, Y* together in same context in the text, and *FX* is occurring number of the word *X* in the text and also *FY* represent the occurs number of the word *Y* in the text.

4.3 ASN model

The following algorithm shown the ASN procedure that applied on all documents and take average of semantic measure for each pair words that occurrence in same topic, has been compared the results of each pair words with standard online word-net by using same similarity metric (Wu & Palmer) as well as the proposal modify in this measurement method, the next example will illustrate how the work of the proposed and the results obtained.

ASN algorithm

```
Input: text after pre-processing (normalization, filtering, stemming)
Output: semantic net represents the relevance relationship between the words of text
Begin
    Get maximum word frequency (wf) from text; Assign word of (wf) as a root of tree;
    While not end of text do Begin
            Get next three word from text (w1, w2, w3);
            Get max word of sequence (w1, w2, w3) and drop word that found in the
        tree;
            Create (N) node(s) where N represent the re mind words (w1, w2, w3);
            Sort nodes from largest to smallest;
            For each remind word (Q) in (N) node do
                    If word Q larger than or equal to largest node in same path then
```

```
                        create node (M) in new path from Root;
                        assign recent largest word from (N) node in (M) node;
                Else
                        create node (K) from node (S) where (S) is larger than (K)
                   in the same path;
                End If
        End For
        Skip to the word which is following max recent Word in text;
    End While
    Return Semantic net;
End.
```

4.4 Simple Example

If the input as the following text to proposed model:

A woman is slicing an onion.
A woman is cutting an orange.
A person is preparing shrimp.
A person is preparing dinner.
A person is cutting broccoli.
A woman is person slicing cucumber

after apply the preprocessing (removing stop words, stemming and tokenization) on the text it will become as follows:

(women, slice, onion, women, cut, orange, person, prepare, shrimp, person, prepare, dinner, person, cut, broccoli, women, person, slice, cucumber).

Now enter this processed text in the ASN model to build a semantic net from this text (fig.3):

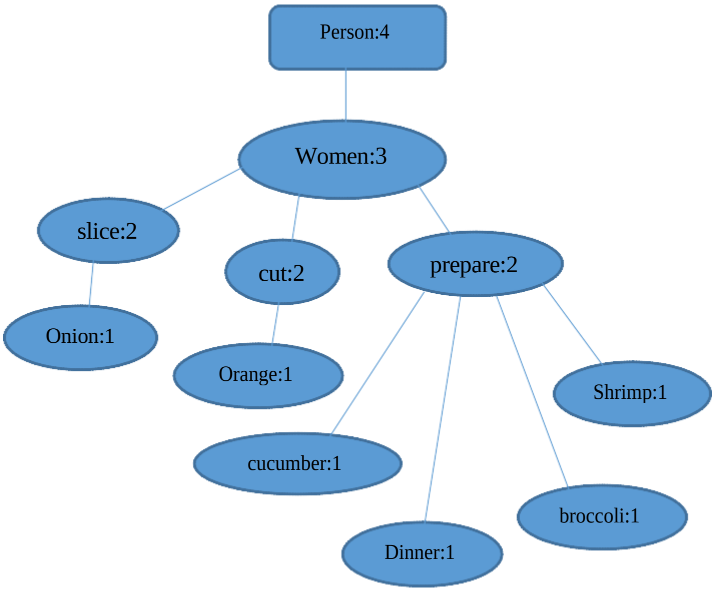


Fig. 3: hierarchal representation of ASN model

The hierarchical structure above appearance the root that contained the word that have largest occurs in the text and others nodes Contains the rest words of the text.

5. Experimental Results

Results collected from (MultiLing Pilot 2011) from https://nlp.stanford.edu/IR-book and https://github.com/trec-kba/many-stop-words/tree/master. Dataset by using fifty documents form three main areas in the dataset: global accidents, weather and sports and apply the ASN model in the section on each document from dataset, After apply this procedure on all documents and take average of semantic measure for each pair words that occurrence in same area, and compared the results of each pair words with standard online word-net by using many of different relatedness and similarity metric measures which depended on semantic net model to estimate the semantic ratio be- tween each pair of words, these compared methods are gloss vector, path length and Wu & Palmer, In addition to the proposal of modify in Wu & Palmer measurement method using ASN model.

Table 1 shown the number of samples for each pair words which are collected from three variant areas in dataset, and compared the semantic relatedness measure which obtained from ASN model for each pair words with number of relatedness measures in standard online English WordNet (WordNet, 2016). All obtained results are normalized by divided on the largest semantic relatedness degree achieved in the same text.

Table 1 : the number of samples for each pair words

No.	Pair words	Gloss vector	Path length	W-P	ASN
1	Tour turn	1.0	1.0	1.0	1.0
2	Tour Team	0.16	0.11	0.428	0.64
3	Tour won	0.18	0.2	0.5	0.6
4	Team Win	0.27	0.2	0.375	0.65
5	Pilot Autopilot	0.53	0.12	0.66	1.0
6	Host Event	0.24	0.14	0.5	0.59
7	Stage Leader	0.26	0.11	0.6	0.74
8	Accident Hospital	0.21	0.09	0.37	0.5
9	Support Hospital	0.15	0.16	0.73	0.85
10	Ocean Ship	0.55	0.07	0.33	0.64
11	Pilot Plane	0.33	0.14	0.60	0.71
12	System Autopilot	0.21	0.25	0.82	0.85
13	Airport Flight	0.31	0.14	0.63	0.70
14	Tsunami Disaster	0.51	0.5	0.94	1.0
15	Death Toll	0.21	0.2	0.95	1.0

The second column in above table appears the pairs of words selected randomly as a samples from dataset which used, others columns compare number of semantic measures from English word-net database and the last column in this table includes the semantic relatedness measure that computed by using the proposed algorithm (ASN), the results above show the relevance degree between pair words by testing of documents in the dataset, and the texts that belongs to same area have been independently tested.

The evaluation of ASN semantic model done by used WordSim353 dataset, it's a specific dataset consist a 353 pairs of words with their semantic relatedness measures, this semantic measures computed based on human judgment and considered as a gold standard to represent the human opinion about the relatedness degrees between the words of this dataset, Table 2 illustrates the Pearson correlation coefficient between the methods of Wordnet (gloss vector, path length, Wu & Palmer) and the ASN semantic model using MultiLing Pilot 2011 by comparing about 125 of words pairs.

Table 2: Pearson correlation coefficients of numbers of semantic measures in online Wordnet with ASN semantic model on using dataset.

No.	Method	Correlation
1	Path length	0.225
2	Gloss	0.258
3	W-P	0.423
4	ASN	0.529

6. Conclusion

In the summary of this proposal illustrate that this method depend on construct an objective relation between words of text , where it can be through a few texts and words find real relevance degree between terms statically and in the order of words within text, but the working of this proposed system must be coupled with the existence of the text talk about a specific topic and harmonic and harmonious in order to achieve the aim of it, so shown that this addition through this proposal make a distributional semantic get a new feature through hierarchical structure that make words with less occurrence in the text to be affiliate for the most frequent words within text, since it gives the weight of the word exactly proportional to its position and importance in that text through compute the average of relatedness degree to each word with others words by summation the this semantic degrees and divided on number of words in the text to obtain the weight of word relative to the rest of the text, from all of the above is the main aim of this proposal lies in supporting the process of text clustering and classification to improving the documents analysis.

References

Xiaofei Zhoua, Yue Hua, Li Guoa, (2014). Text Categorization Based on Clustering Feature Selection, 2nd International Conference on Information Technology and Quantitative Management, ITQM, China.

Tao Liu ZhengChen, Benyu Zhang, Wei-yingMa, Gongyi Wu Nankai, (2004). Improving Text Classification using Local Latent Semantic Indexing, Microsoft Research Asia Nankai University, China, Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04).

Wael H. Gomma and Aly A. Fahmy, (2012). short answer grading using string similarity and corpus-based similarity, international journal of advanced computer science and applications, Vol. 3, No. 11.

Wael H. Gomaa and Aly A.Fahmy, (2013). A survey of text similarity approaches, international journal of computer applications, Vol. 68,No. 13.

Marco Baroni and Alessandro Lenci, (2010). Distributional Memory: A general framework for corpus-based semantics, Computational Linguistics, Vol. 36, No 4.

Peter Turney and Patrick Pantel, (2010). From frequency to meaning: Vector space models of semantics, Journal of Artificial Intelligence Research 37: 141–188.

Bamshad Mobasher, Olfa Nasraoui, Bing Liu and Brij Masand, (2004). Advances in Web Mining and Web Usage Analysis, 6th International Workshop on Knowledge Discovery on the Web, USA.

Simmons and Robert, (1978). Semantic Net, Proceedings of International Computer Symposium, Nanking, Taipei, Vol. 1, December 18-20, China.

Wu Z. and Palmer M. (1994). Verbs semantics and lexical selection, Proceedings of the 32nd Meeting of Association of Computational Linguistics.

WordNet version 3.0, (2016). http://maraca.d.umn.edu.

Tao Liu ZhengChen, Benyu Zhang, Weiying Ma and Gongyi Wu, (2010). Scalability Issues in Authorship Attribution, PhD Thesis, School of Computer Science and Information Technology, Dutch UPA University.

Evgeniy Gabrilovich and Shaul Markovitch , (2006). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis", Department of Computer Science Technion-Israel Institute of Technology.

Peng Wang, Jiaming Xu, Bo Xu, Cheng-Lin Liu, Heng Zhang Fangyuan Wang, Hongwei Hao, (2015). Semantic Clustering and Convolutional Neural Network for Short Text Categorization, the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 352–357, Beijing, China, July 26-31.

Paola Velardi, Roberto Navigli and Stefano Faralli, (2011). Onto Learn Reloaded: A Graph-Based Algorithm for Taxonomy Induction, Computational Linguistics Vol. 39, No.

3.

Xinghua Lu, Bin Zheng, Atulya, Velivelli and ChengxIiang Zhai,(2017). A framework for enriching lexical semantic resources with distributional semantics “, Natural Language Engineering 1, Printed in the Cam- bridge University Press, United Kingdom.

Amal Bouraoui, Salma Jamoussi and Abdelmajid Ben Hamadou, (2017), A New Method for the Construction of Evolving Embedded Representations of Words, IEEE/ACS 14th International Conference on Computer Systems and Applications.

Douwe Kiela and Stephen Clark, (2014). A Systematic Study of Semantic Vector Space Model parameters, In Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality.

Zellig Sabbettai Harris, (1954). Distributional Structure, Word, Vol.10, No 2-3, p.p:146-162.

Patrick Henry Winston, (1992). “Artificial Intelligence”, Third Edition, Pearson Education.

Canada & Moeiz Miraoui, (2016). A modification of Wu and Palmer Semantic Similarity Measure, technical report, Dept. of Electrical Engineering, University of Gafsa, Tunisia.