Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

NOUN PHRASE BASED WEGHTING SCHEME FOR SENTENCE SIMILARITY MEASUREMENT

A. T. Mahmood[1,2,*], S. S. Kamaruddin[1] and R. K. Naser[1,3]

[1]School of Computing, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
[2]Technical Instructors Training Institute, Middle Technical University, Ministry of Higher Education and Scientific Research, Iraq
[3]Military Training Directorate, Ministry of Defense, Iraq

Published online: 17 October 2017

ABSTRACT

The need for an effective text similarity measures has led many previous studies to propose different text weighting schemes to enhance the overall performance of sentence similarity measures. Term Frequency Inverse Document Frequency (TF-IDF) is a weighting method that is commonly used to determine the occurrence of a term/word in a document. This paper proposes the use of Noun Phrase (NP) based TF-IDF weighting scheme in order to empower the efficiency of text similarity. A total of eight pairs of statements were used to validate the proposed method. The obtained results were then compared with the standard TF-IDF weighting scheme. The result shows that NP-TF-IDF has significantly improved the performance of text similarity measures as compared to the standard TF-IDF. Our findings may offer the necessary insights related to the development of text similarity applications.

Keywords: TF-IDF; noun phrase chunking; sentence similarity.

Author Correspondence, e-mail: ayhar1973@gmail.com

doi: http://dx.doi.org/10.4314/jfas.v9i5s.11

1. INTRODUCTION

Text similarity measures have been widely used in several natural language processing applications such as automatic essay grading, paraphrase recognition, etc[1-3]. Previous studies on text similarity were mostly concerned about the semantic typing in terms of two mechanisms: the detection of similarity and difference in the form of judgments of likeness in which other potential inconsistency that can be resulted from judgments of difference. The current applications of text similarity operates as a naturally skewed or artificially imbalanced[4-5].With this in mind, weighting schemes are used to promote the process of text similarity. Term Frequency Inverse Document Frequency (TF-IDF) is an example of weighting scheme that help to maximize the performance gain. It is mainly used to help in classifying the indexing terms by assigning them weights corresponding to how well they are in improving both the recall and precision of the retrieval [6]. A number of efforts were carried out to improve the performance of current weighting schemes [7] to which some limitations may be resulted due to the use of probabilistic model to describe a set of possible probability distributions for a set of observed data. It can also be used to determine the distribution in the probabilistic model [8-9].This paper attempts to determine the potential of Noun Phrase (NP) in text similarity. This paper discusses the current application of TF-IDF and compare its performance with NP-TF-IDF. Our results showed the potential of NP-TF-IDF as compared to TF-IDF alone. The paper is organized as follows: Section 2 addresses the related works followed by discussion on the proposed method in section 3. The results is presented in section 4 and the paper ends with discussion and conclusion in section 5.

2. RELATED WORKS

Text similarity is the measure used to determine the key point in text categorization that is used to integrate the meaning of texts into the similarity [10].

Table 1 summarizes the types of text similarity measures that was classified by [11]and used by previous studies. The summarization was based on the aggregation that was mostly found in bidirectional, cosine and unidirectional aggregates. In this study, the cosine aggregate is used in which the shortest path was found to be used for measuring word similarity.

Table 1. Types of text similarity measures

Reference	Word Similarity Measures	Aggregation
Mihalcea, Corley [12]	PMI-IR [13], Latent Semantic Analysis [14], additional measures [15], Shortest Path [16], Explicit Semantic Analysis [17], Corpus-based word similarity [18]	Bidirectional aggregate
Tsatsaronis, Varlamis [19]	Shortest Path (16)	Bidirectional aggregate
Li, McLean [20]	Shortest Path (16)	Cosine aggregate
Ho, Murad [21]	Longest Common Subsequence [22]	Unidirectional aggregate
Islam, Milios [10]	Word trigrams (Web1T)	Unidirectional aggregate

The cosine similarity is mostly performed on the aggregation of two vectors. Its main purpose is to determine the cosine value of the angle between these two vectors with regards to certain features that may present in a document. It is also used to calculate the orientation of a document on a normalized space[23]. The process associated with calculating the cosine similarity is described in the following equation for the $cos\theta$:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \qquad (1)$$

whereAi and Bi are components of vector A and B respectively.

Determining the similarity of certain words in a document is one of the main concerns of researchers nowadays. Previous studies recommended the use of TF-IDF [5] as a method for processing word into separate vectors. This can be illustrated by:

$x = (x_1, x_2,...,x_n)$(2)

where$x$ refers to the text record.

A. TF-IDF

TF-IDF is a weighting method used to determine the occurrence of a term/word in a document.

It is also used for weighting certain terms for information retrieval and text mining. The value associated with the weighting of term's importance is based on the increases in the number of frequency to the number of times a word appears in the document [24]. It is calculated as follow:

$tfidf\ (t,d) = tf(t) * idf(t)$ (3)

In which tf value is calculated based on the following equation:

$tf(t,d) = f(t)/\ n$ (4)

where$tf(t,d)$ refers to the term frequency,n refers to the total number of terms; d refers to the supplied dataset and $f(t)$ refers to the term frequency. However, the IDF part will be also calculated in order to determine the reduced weight for high frequency terms in remaining documents. This will be obtained by using the following:

$idf(t) = \log |D| \ / \ |\{d : t \in d\}s|$ (5)

where$D$ refers to the total number of documents, d is the quantity of documents and $idf(t)$ refers to the inverse terms.

There are many alternatives for TF-IDF weighting schemes, for example:

2.1. Robertson and Jones (RSJ) Weighting Scheme

RSJ weight by [25]. It is calculated based on the following:

$$\mathrm{id}f_i = \log \frac{|D|}{n_i}, \qquad (6)$$

where$D$ refers to the total number of documents used for the training, while $n_i$refers to the total number of documents which have the *ith* word.

2.2. BM25 Weighting Scheme

BM25 by Robertson, Walker [26] and it is calculated using the following equation:

$$\mathrm{id}f_i = \frac{(k_1 + 1) \cdot \mathrm{t}f_i}{K + \mathrm{t}f_i} \log \frac{|D|}{n_i}, \qquad (7)$$

where$k1$ and K refers to the standard parameters supplied by the machine in the training phase.

2.3. Term Weight Inverse Document Frequency (TW-IDF) Weighting Scheme

TW-IDF by[27] was also proposed as a weighting scheme that offers graph-of-word model to present any potential relationships between the terms using an un-weighted directed graph of terms. It is calculated as follow:

$$\text{TW-IDF}(t,d) = \frac{\text{tw}(t,d)}{1 - b + b \times \frac{|d|}{\text{avdl}}} \times \log \frac{N+1}{\text{df}(t)}, \qquad (8)$$

where*tw(t,d)* refers to the total weight of the vertex *t* (term *t*) that is presented in the graph-of-word of the document *d*, *b* is a parameter, *N* refers to the total number of documents while *df(t)* refers to the frequency of term *t* for all the documents. In addition, the term *avdl* refers to the document length and *Idl* is the length of the document *d*.

2.4. The Term Depth Distribution Weighting Scheme

The term depth distribution by[28] was also proposed and it is calculated based on the following:

$$w_{ik} = c^{-(f_{ik}-f_{ia})^2} \cdot \log \left[\frac{N \cdot D_k}{d_k \cdot L_k} \right], \qquad (9)$$

where f_{ia} refers to the average value of the frequency in a document *I*, *fik* represents the frequency of the term *k,* L_k represents the accuracy of term k appeared in the corpus, D_k refers to the terms in documents, where $w_{ik}$is the term significance of term kin document *i.*

3. METHODOLOGY

This study compared the similarity of sentences using noun phrase based TF-IDF. Text character feature types were compared in terms of the NP. The comparison was mainly conducted to show the level of text similarity between different sentences. The dataset used in this experiment are obtained from sample datasets in [33].Specifically, we used a noun phrases filter in order to determine the keyphrases within the selected dataset. This is because the noun phrase filter contain different small set of patterns; it also helps to ensure proper extraction. The following are the main steps used in this study in order to calculate the NP TF-IDF.

1. Noun+Noun
2. (Adj|Noun)+Noun
3. ((Adj|Noun)+((Adj|Noun)∗(NounPrep)?)(Adj|Noun)∗)Noun

Fig. 1 shows the major phases used in this study. The process starts with collecting 8 pairs of sample sentences as the main data for this study followed by the preprocessing phase in which the collected data will be cleaned for further use. In the preprocessing phase, the data was transferred into a suitable form, to be used later in the relevant machine learning algorithms. Then, the inserted dataset was represented in the form of extracted features along with their associated weights. Here, we considered the main aspects related to Tokenisation to construct a single word from the processed sentence using Natural Language Toolkit (NLTK). Then, Stopword removal is implemented to eliminate the non-meaningful words from the sentences. Finally, Lemmatization and stemming were used to manage any potential impact of grammatical form in the context.

NP chunking is utilized on the processed data in order to separate the words in each sentence so that TF-IDF technique can be applied on. Then, the chunked text is classified and evaluated using performance measures of precision, recall and F-measure or F1. Further explanation about each phase is provided in the following sections.

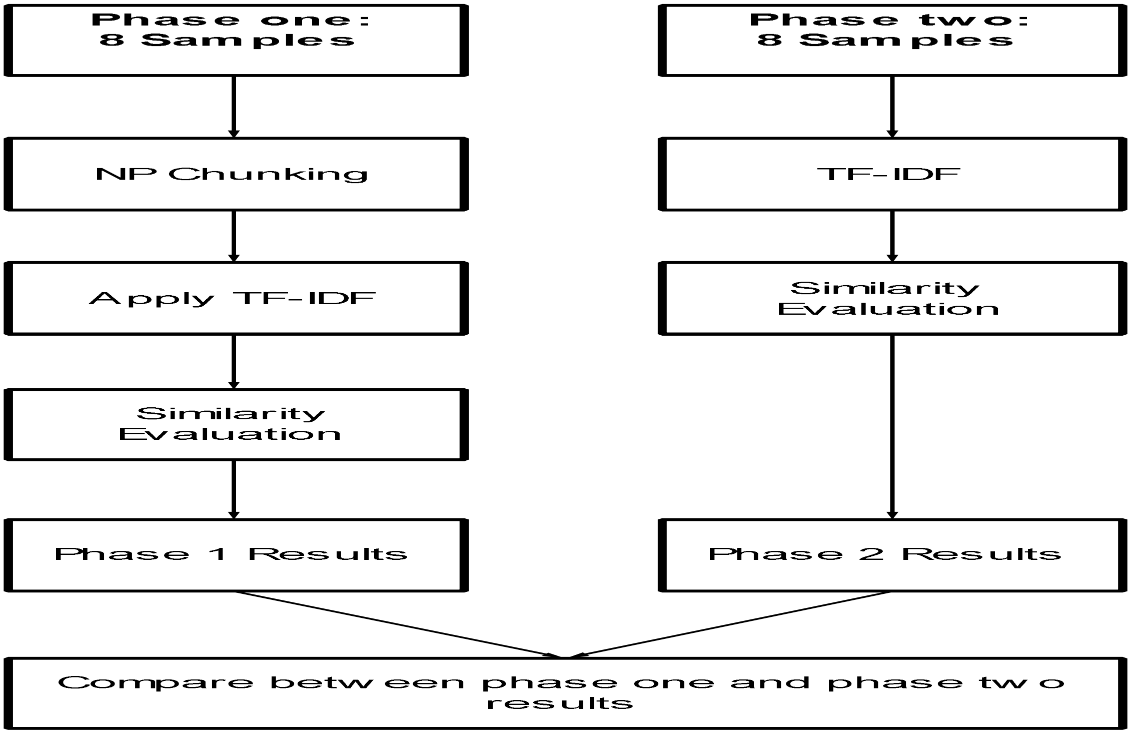

Fig.1. The experimental settings

3.1. Procedure

The main steps followed in this paper are divided into two phases in which eight text samples were used in each phase. Phase one consists of using NP chunking along with TF-IDF while phase 2 consists of using TF-IDF alone. Then, the similarity for each phase was obtained and compared.

3.2. Data Set

In the initial experiment, the research considered the unstructured text datasets from corpus used in the work of [33].We have randomly selected four positive and four negative samples and their corresponding similarities from 8 pairs sample sentences to demonstrate the feasibility and validity of the proposed method. Table 2 shows the selected samples.

Table 2. Eight samples for the initial experiment

Sample No.	Samples
Sample1:	“Taha is married to former Iraqi oil minister Amir Muhammed Rasheed, who surrendered to U.S. forces on April 28.” “Taha’s husband, former oil minister Amer Mohammed Rashid, surrendered to U.S. forces on April28.”
Sample2:	“On July 22, Moore announced he would appeal the case directly to the U.S. Supreme Court.” “Moore of Alabama says he will appeal his case to the nation’s highest court.”
Sample3:	“Six Democrats are vying to succeed Jacques and have qualified for the Feb. 3 primary ballot.” “Six Democrats and two Republicans are running for her seat and have qualified for the Feb. 3 primary ballot.”
Sample4:	“Agriculture Secretary Luis Lorenzo told Reuters there was no damage to the vital rice crop as harvesting had just finished.” “Agriculture Secretary Luis Lorenzo said there was no damage to the vital rice crop as the harvest had ended.”
Sample5:	“A soldier was killed Monday and another wounded when their convoy was ambushed in northern Iraq.” “On Sunday, a U.S. soldier was killed and another injured when a munitions dump they were guarding exploded in southern Iraq.”
Sample6:	“Perkins will travel to Lawrence today and meet with Kansas Chancellor Robert Hemenway.”“Perkins and Kansas Chancellor Robert Hemenway declined

	comment Sunday night.”
Sample7:	“‘I am proud that I stood against Richard Nixon, not with him,’ Kerry said.” “‘I marched in the streets against Richard Nixon and the Vietnam War,’ she said.”
Sample8:	“The report by the independent expert committee aims to dissipate any suspicion about the Hong Kong government’s handling of the SARS crisis.” “A long awaited report on the Hong Kong government’s handling of the SARS outbreak has been released.”

4. RESULTS AND DISCUSSION

Table 3 presents the text similarity found in the supplied statements for both methods. Our results revealed that the NP based TF-IDF offers a better similarity relatedness with human evaluation than TF-IDF alone. This indicates the potential of using NP TF-IDF for text similarity.

Table 3. Evaluation results for the comparison

	Text Similarity (Human Evaluation)	**TFIDF**	**NP TFIDF**
Sample: 1	1	0.58	0.59
Sample: 2	1	0.41	0.62
Sample: 3	1	0.72	0.66
Sample: 4	1	0.78	0.91
Sample: 5	0	0.43	0.13
Sample: 6	0	0.52	0.62
Sample: 7	0	0.37	0.0
Sample: 8	0	0.58	0.66

The visual comparison of the text similarity between TF-IDF and NP TF-IDF is shown in Fig. 2, which illustrated that NP TF-IDF offer better performance when comparing it with the standard TF-IDF. However, there is still a need for extra efforts in order to validate the feasibility of the proposed method in larger data group.

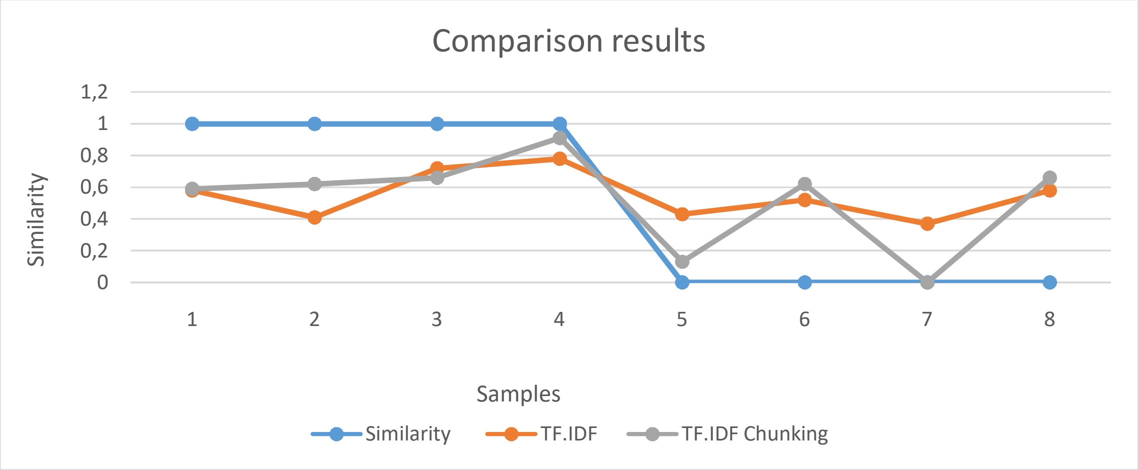

Fig.2. Comparison of text similarity between TF-IDF and NP TF-IDF

The result supported our claim about the potential of NP based TF-IDF. In addition, the proposed scheme offer better results by determining the relevance of text in document queries using TF-IDF weighting scheme. Our results can help provide the necessary evidence about the potential of NP for TF-IDF.

5. CONCLUSION

This study introduced the potential of using NP in TF-IDF for text similarity. Such use was motivated by the lack of studies to document NP effectiveness in text similarity measures. Our initial findings on eight samples appear to validate the potential of NP TF-IDF in text similarity measures. It also offers the necessary evidence about the effectiveness of using NP for managing the set of nouns in a text. This study supports previous efforts by[29]who calculated the semantic similarities between sentences and performed a comparative analysis among identified similarity measurement techniques. It also enrich the study by [30] who wondered the role of NP in categorizing subjective sentences from objective sentences. It also adds to the previous effort by[31] who aimed at increasing the accuracy of text mining tasks with emphasis on concept extraction from text in concept-level text analysis. In addition, the finding of this study is inline with other previous works who addressed the potential of NP for various text similarity measures. For example, in [32]recommended the generalization of two noun phrases by extending the mechanism of logical generalization towards syntactic parse trees and attempt to detect weak semantic signals from them.

Despite the effectiveness of NP TF-IDF, some limitations still persist. For example, our work was limited to relatively small sample of text to validate the proposed NP TF-IDF. In addition, the analysis method was limited to the examination of text similarity through human evaluation which may not necessarily provide an in-depth understanding of NP TF-IDF feasibility. Therefore, future works can be conducted to implement the proposed method for text classification using larger data set and perform comparison with other weighting schemes.

6. REFERENCES

[1] Gomaa WH, Fahmy AA. A survey of text similarity approaches. International Journal of Computer Applications, 2013, 68(13):13-18

[2] Buscaldi D, Le Roux J, Flores JJ, Popescu A. Lipn-core: Semantic text similarity using n-grams, wordnet, syntactic analysis, esa and information retrieval based features. In2nd Joint Conference on Lexical and Computational Semantics, 2013, pp. 1-7

[3] Croft, D, Coupland S, Shell J, Brown S.A fast and efficient semantic short text similarity metric.In 13th UK Workshop Computational Intelligence, 2013, pp. 221-227

[4] Liu Y, Loh HT, Sun A. Imbalanced text classification: A term weighting approach. Expert Systems with Applications, 2009, 36(1):690-701

[5] Ramos J. Using TF-IDF to determine word relevance in document queries. In1st Instructional Conference on Machine Learning, 2003, pp. 1-4

[6] Sun YH, He PL, Chen ZG. An improved term weighting scheme for vector space model. In IEEE International Conference on Machine Learning and Cybernetics, 2004, pp. 1692-1695

[7] He N, Xu D, Zhu Y, Zhang J, Shen G, Zhang Y, Ma J, Liu C. Weighted average current control in a three-phase grid inverter with an LCL filter. IEEE Transactions on Power Electronics, 2013, 28(6):2785-2797

[8] Blei DM. Probabilistic topic models. Communications of the ACM, 2012, 55(4):77-84

[9] Hong TP, Lin CW, Yang KT, Wang SL. Using TF-IDF to hide sensitive itemsets. Applied Intelligence, 2013, 38(4):502-510

[10]Islam A, Milios E, Kešelj V. Text similarity using google tri-grams. InCanadian

Conference on Artificial Intelligence, 2012, pp. 312-317

[11] Bär D, Zesch T, Gurevych I. Composing measures for computing text similarity.Technical Report TUD-CS-2015-0017, Hessen, TechnischeUniversität Darmstadt, 2015

[12]Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity. In 21st National Conference on Artificial Intelligence, 2006, pp. 775-780

[13]Thede SM. Parsing and tagging sentences containing lexically ambiguous and unknown tokens. Phd thesis, Indiana: Purdue University, 1999

[14]Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse Processes,1998, 25(2-3):259-284

[15]Mohler M, Mihalcea R. Text-to-text semantic similarity for automatic short answer grading. In12th Conference of the European Chapter of the Association for Computational Linguistics, 2009, pp. 567-575

[16]Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 1989,19(1):17-30

[17]Gabrilovich E, Markovitch S. Computing semantic relatedness using wikipedia-based explicit semantic analysis. InInternational Joint Conference on Artificial Intelligence,2007, pp. 1606-1611

[18]Islam A, Inkpen D. Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data, 2008, 2(2):1-25

[19]Tsatsaronis G, Varlamis I, Vazirgiannis M. Text relatedness based on a word thesaurus. Journal of Artificial Intelligence Research, 2010, 37(1):1-40

[20]Li Y, McLean D, Bandar ZA, O'shea JD, Crockett K. Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(8):1138-1150

[21]Ho C, Murad MA, Kadir RA, Doraisamy SC. Word sense disambiguation-based sentence similarity. In23rd International Conference on Computational Linguistics, 2010, pp. 418-426

[22]Yang D, Powers DM. Measuring semantic similarity in the taxonomy of WordNet. In 28th Australasian conference on Computer Science, 2005, pp. 315-322

[23]Sidorov G, Gelbukh A, Gómez A H, Pinto D. Soft similarity and soft cosine measure:

Similarity of features in vector space model. Computación y Sistemas, 2014,18(3):491-504

[24]Aizawa A. An information-theoretic perspective of tf–idf measures. Information Processing andManagement, 2003,39(1):45-65

[25]Robertson SE, Jones KS. Relevance weighting of search terms. Journal of the American Society for Information Science,1976, 27(3):129-146

[26]Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M. Okapi at TREC-3. In D. K. Harman (Ed.), Overview of the 3rd Text REtrieval Conference (TREC-3)computer systems technologyissue 500, part 225 of NIST special publicationNIST special publication: Computer systems technology, Pennsylvania, Diane Publishing, 1995, pp. 109-126

[27]Rousseau F, Vazirgiannis M. Graph-of-word and TW-IDF: New approach to ad hoc IR. In22nd ACM international conference on Information and Knowledge Management, 2013, pp. 59-68

[28]Zhang J, Nguyen TN. A new term significance weighting approach. Journal of Intelligent Information Systems, 2005, 24(1):61-85

[29]Saad SM, Kamarudin SS. Comparative analysis of similarity measures for sentence level semantic measurement of text. In IEEE International Conference on Control System, Computing and Engineering, 2013, pp. 90-94

[30]Riloff E, Wiebe J, Wilson T. Learning subjective nouns using extraction pattern bootstrapping. In7th Conference on Natural Language Learning, 2003, pp. 25-32

[31]Poria S, Agarwal B, Gelbukh A, Hussain A, Howard N. Dependency-based semantic parsing for concept-level text analysis. InInternational Conference on Intelligent Text Processing and Computational Linguistics, 2014, pp. 113-127.

[32]Galitsky B. Machine learning of syntactic parse trees for search and classification of text. Engineering Applications of Artificial Intelligence, 2013, 26(3):1072-1091

[33] Li L,Hu X,Hu B,Wang J,Zhou Y. Measuring sentence similarity from different aspects. In 8th International Conference on Machine Learning and Cybernetices, 2009, pp. 2244-2249

How to cite this article:
Mahmood A T, Kamaruddin S S, Naser R K. Noun phrase based weghting scheme for sentence similarity measurement. J. Fundam. Appl. Sci., 2017, *9(5S), 131-143.*